A Deep Learning-Based Method for Drivers' Shoe-Wearing Recognition

Baoyue Hu(✉) and Xing Hu

University of Shanghai for Science and Technology, Shanghai, China
fanthal@163.com

Abstract. What types of shoes that the driver should be wear on driving is under the clear regulation. Non-standard shoe-wearing such as wearing high-heeled shoes, platform shoes, slippers or with bare feet will bring great safety risks and may lead to traffic accidents. According to statistics by traffic department, many traffic accidents are caused by irregular shoe-wearing. Although computer vision has been applied for monitoring driver's behavior in many aspects, such as driver's face, eye, and hand, there is still no computer vision-based method or hardware device on the vehicle that to monitor the driver's shoe-wearing. Therefore, if the driver's illegal shoe-wearing behavior can be identified before driving, it can play an important role in reducing the incidence of traffic accident. The main difficulties in drivers' shoe-wearing detection lie in the diversity of shoe types, the variety of foot postures, and the complexity of the detection environment. Therefore, the traditional computer vision method will be high error detection rate in such scenes. In this paper, a deep learning-based method for detecting abnormal shoe-wearing of drivers before driving is proposed. Different models such as SVM, Fast-RCNN, Faster-RCNN, YOLO v2, YOLO v3, YOLO v4 are used to identified drivers' shoe-wearing. To the best of our knowledge this is the first work that using the computer vision technology for automatic monitoring on drivers' shoe-wearing. The experimental results show the proposed method can identify the drivers' shoe wearing effectively and efficiently. It has high application value for improving traffic safety.

Keywords: YOLO v4 · Deep Learning · Computer Vision Technology · Shoe-Wearing Recognition

1 Introduction

1.1 A Subsection Sample

How drivers wear shoes while driving is clearly specified in traffic laws. Some irregular shoe-wearing behaviors, such as wearing high heels, slippers, platform shoes, and even bare feet, are not allowed while driving. Because in this shoe-wearing case, the driver often fails to jam on the brakes timely and effectively in the event of an emergency, which leads to traffic accidents. According to statistics, 60% of traffic accidents every

H. R. Karimi and N. Wang (Eds.): S-Cube 2022, LNICST 487, pp. 167–178, 2023.
https://doi.org/10.1007/978-3-031-34899-0_11

year are related to the drivers' irregular shoe-wearing [1]. However, the supervision of drivers' abnormal shoe-wearing is not as effective as other violations. The traffic cameras cannot monitor it effectively because the driver's feet are inside the cab. Even if the traffic police conduct an investigation of the vehicle, they will not be able to find the illegal shoe-wearing behavior if the driver does not get out of the car. What's more, it is difficult for the limited number of traffic police to monitor the large number of vehicles effectively.

With the development of computer vision and artificial intelligence technology, more and more supervision of drivers' violations is to use computer technology to achieve intelligence and automation. For example, not wearing a seat belt [2], holding a mobile phone with hands while driving [3], etc. can be detected by cameras installed above the road. However, since the drivers' feet are inside the cab, it is difficult to detect by the camera. Even the surveillance cameras installed inside the cab only focus on monitoring the drivers' face and hands. So far, there is no relevant in-vehicle equipment specifically used for monitoring the drivers' shoes. However, unlike shoe types identification in ordinary situations, the main difficulties in identifying the drivers' shoe-wearing behavior lies in the following points:

1) Diversity of the driver's foot posture.
2) Variety of driver's shoe types.
3) Light and occlusion problems in the cab. The existence of the above problems makes the traditional image processing and machine learning methods usually have a high recognition error rate.

Fig. 1. Examples of Different Types of Shoes

This paper proposes a computer vision-based method of detection drivers' abnormal shoe-wearing to deal with these problems. The method proposed in this paper divides the drivers' shoe-wearing into two categories (as shown in Fig. 1): The first category is the types of shoes suitable for drivers to drive, including flat shoes, cloth shoes, leather shoes, etc. The other category is the types of shoes that are not suitable for driving, including slippers, high heels, platform shoes and bare feet, etc. The method captures

pictures of the drivers' feet as they enter the cab with tiny cameras mounted inside the car doors. After that, use the captured pictures to train the deep neural network model, and output two types of labels which include legal shoe-wearing and illegal shoe-wearing. When it is detected that the type of shoes is illegal, the vehicle will be restricted from starting through the vehicle control system, and then the driver will be prompted to change to legal shoes by voice. This paper trains six neural network models of SVM, Fast-RCNN, Faster-RCNN, YOLO v2, YOLO v3 and YOLO v4 respectively. The results of the experiment show that the YOLO V4 model has the highest recognition rate. Meanwhile, it can meet the real-time requirements. To our best knowledge, this is the first work to automatically identify whether the driver's shoe-wearing is legal.

The main contributions of the paper are as follows:

1) We use the real data set to train the learning model and come to the conclusion that YOLO V4 model has the best performance in detection drivers' shoe-wearing behavior.
2) We construct a dataset for the identification of driver's abnormal shoe-wearing for the first time. The dataset contains a large number of legal and illegal shoe-wearing samples from real-world, which can be used for training deep learning models to get the most accurate conclusion.
3) The model proposed in this paper provides an effective method for the detection of drivers' shoe-wearing behavior in daily life, which can greatly reduce the incidence of traffic accidents.

This paper is organized as follows: The second paragraph describes the application of computer vision in driver-specific monitoring. The third paragraph describes the hardware of the proposed method in this paper and the flow chart of the algorithm. The fourth paragraph is the specific experimental content, which verifies the effectiveness of the method in this paper. The fifth paragraph concludes the paper.

2 Pertinent Literature

Since this paper propose a computer vision-based recognition method for driver's abnormal shoe-wearing for the first time, there is no similar related literature at present. However, there are many ways based on computer vision can be used to monitor drivers. This paper will review the related methods of driver's face and hand monitoring based on computer vision.

2.1 Computer Vision-Based Driver's Face Monitoring Method

Similar to driver's shoe recognition, in unconstrained face recognition, the face images may have many variations, such as low resolution, pose variation, complex light and motion blur, which will result the low recognition accuracy. Traditional face recognition algorithms, such as the Eigenfaces [4], Bayesian [5], support vector machine (SVM) based [6] can assist computers to complete basic face recognition. But for the more general case, such as unconstrained face matching, the traditional algorithms may not do well. As deep learning models exhibit the superior accuracy and robustness in extracting

features, significant progress has been made in face recognition with the development of deep learning technology. Among the deep learning models, the convolutional neural network (CNN) has become the most popular one due to its excellent performance. Before the method proposed in this paper, deep learning models such as CNN [7] have been used for driver's face detection [8]. The driver's face detection often locates the face first, and the localization work is often indexed by the eyes or mouth [9]. If the accuracy of face recognition and classification needs to be guaranteed, it is necessary to ensure that the facial features of each person are extracted effectively as much as possible. The deep learning models provide us with ideas to solve these problems. Experiments show that the deep learning models can still extract the corresponding facial features well in complex situations, such as makeup considered eye status recognition for driver drowsiness [10], the anti-occlusion face recognition for drivers [11].

2.2 Computer Vision-Based Driver's Hands Monitoring Method

Different from the driver's facial recognition, the difficulty in monitoring driver's hands movements mainly lies in the diversity of postures. For this problem, it is often necessary to set a dataset with a sufficiently large sample size to solve it [12]. Because of the lack of driver's hand features, the classification of driver's behavior mainly relies on the large amount of existing data for training. The superior data processing ability of the deep learning models also makes it perform suitably in this problem [13, 17]. Before deep learning-based computer vision technology was used for driver behavior detection, methods for recognition and detection of different actions have been proposed. For example, RGB camera-based fallen person detection system was proposed with the emerging computer vision recognition algorithm YOLO [14]. The current mainstream method for driver's behavior detection is to pre-train the neural network through a dataset composed of a large number of samples, and then perform backpropagation and optimize the dataset according to the new images that are continuously captured. However, existing techniques utilizing image-feature-based for encoding such activity can sometimes misclassify crucial scenarios. The video-feature-based extraction was proposed then [15, 18]. Some efficient and accurate deep learning models like such as YOLO begin become popular and replace the traditional deep learning models like SVM, RCNN [16, 19].

3 Experimental Principles

3.1 System Structure

In this paper, the proposed deep learning-based method of detecting drivers' abnormal wearing of shoes relies on the assistance of miniature cameras and vehicle-mounted miniature computers. Its specific structure is shown in Fig. 2.

As shown in Fig. 2, when a driver is in the driver's seat, the miniature camera collects the images of the driver's feet and transmits the captured real-time images to the vehicle-mounted miniature computer. The vehicle-mounted miniature computer has already been equipped with a neural network that has been trained by a large number of samples. After receiving the real-time images, the vehicle-mounted miniature computer

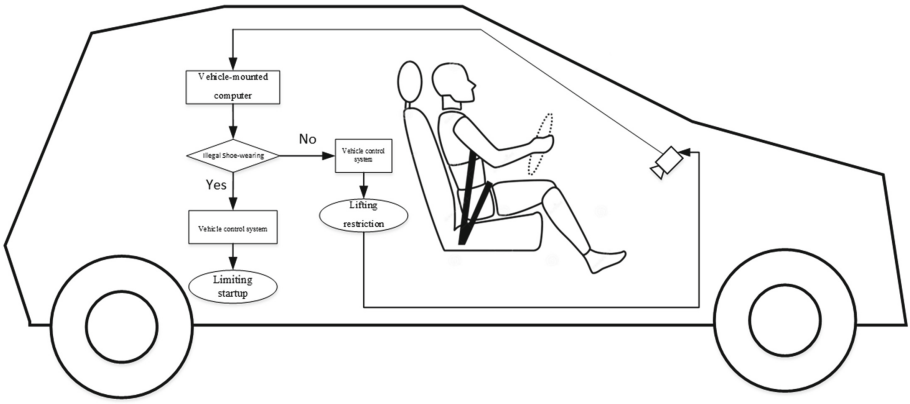

Fig. 2. Structure Diagram of Driver Abnormal Shoe Wearing Detection System

immediately starts to detect, identify and classify such images, and judges whether the driver's shoes are in line with the safe driving norms according to the detection results: if the driver is wearing high-heeled shoes, platform shoes, slippers or barefoot, the vehicle-mounted miniature computer will transmit a signal to the vehicle control system to limit the start of the car and issue an alarm to prevent the driver from driving wearing shoes with potential safety hazards; If the driver is wearing shoes such as flat shoes in line with the safe driving norms, the vehicle-mounted miniature computer will transmit an enabling signal to the vehicle control system to lift the startup restriction, so the driver can drive normally on the road.

3.2 Deep Learning Model

To describe the deep learning-based method of detecting drivers' abnormal wearing of shoes proposed in this paper more clearly, the network training algorithm, model, and structure used in the method are expounded here.

The proposed method is based on the YOLO V4 algorithm. Its main network architecture is shown in Fig. 3.

As shown in Fig. 3, the image standard input size of the network is 608 * 608, so the algorithm will preprocess the images before they enter the baseline network to scale their size to the standard size and uniformize them. In addition, Mosaic data enhancement, CmBN, and SAT [21] will also be performed on the images. To detect the abnormal shoe-wearing behaviors of drivers, the trained network is required to have high precision and high robustness. Therefore, sample data sets are processed by Mosaic for data enhancement, and sample images are randomly scaled, distributed, and spliced, which greatly improves detection speed and reduces memory requirements.

YOLO V4's backbone network (Backbone) is an improvement of YOLO V3's Darknet53 network. Based on the network structure of YOLO V3, five CSP modules are added to the network of the YOLO V4 algorithm. In the course of processing input images and outputting feature maps, traditional computer vision detection methods usually use the DenseNet algorithm. Although this algorithm can complete the tasks of backpropagation

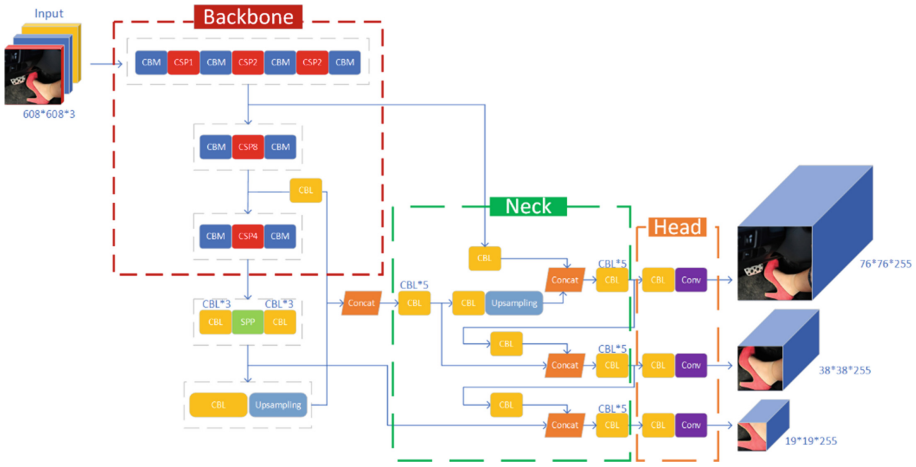

Fig. 3. YOLO v4 Network Architecture

and weight optimization, it will produce a lot of double counting during backpropagation, resulting in heavy workload and slow propagation and optimization speed. The CSP module divides the features of a basic image layer into two parts. One part is processed by the local Dense module and connected to the other part through a cross-stage hierarchical structure, which not only reduces the amount of computation but also ensures the accuracy of the overall structure.

Meanwhile, YOLO V4 uses the Mish activation function (as shown in Formula 1 below):

$$Mish = x * tan\ h(ln(1 + e^x)) \tag{1}$$

The image of the Mish activation function is shown in Fig. 4:

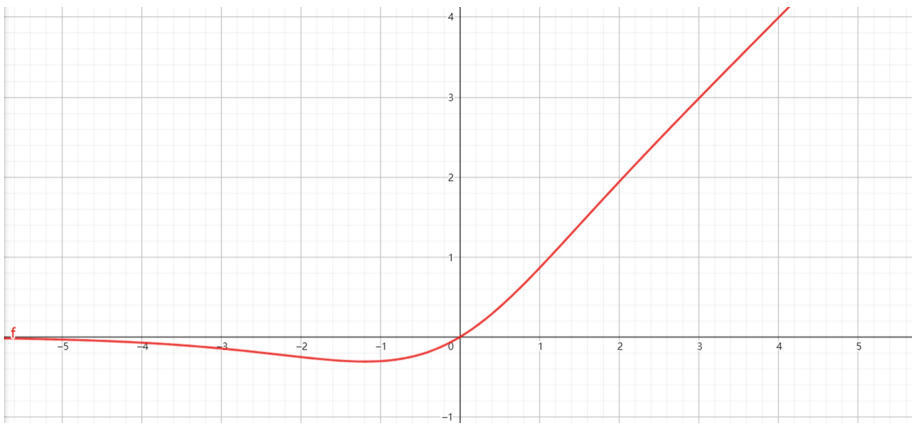

Fig. 4. Image of Mish Activation Function

Compared to the activation function of the previous version of the YOLO algorithm, the Mish activation function has a smoother curve and allows a smaller negative gradient, which ensures the integrity of feature information to a certain extent and has a better gradient descent effect. In addition, the linear function x and logarithmic function ln ensure that the gradient of tanh function doesn't approach 1 as it approaches positive or negative infinity, thus avoiding the saturation problem.

In the Neck part of the network architecture, YOLO V4 adopts SPP [22, 23] and FPN+PAN modules. In the Backbone part, YOLO V4 completes the preliminary shallow feature extraction before turning to the Neck part of the network for feature enhancement. The SPP module adopts the max-pooling mode of 1×1, 5×5, 9×9, and 13×13, and realizes Concat fusion and splicing of feature maps of different scales. The SPP module can greatly improve the receptive field, separating out contextual features while maintaining speed. Domain-specific fine-tuning is then performed to significantly improve performance [24]. In addition, FPN captures strong semantic features and PAN conveys strong localization features, which can accomplish target localization more effectively [25]. As the images of drivers' shoes that may be different in type are captured by the camera at different rays of light and different angles, strong features are required to ensure detection and identification.

As for the detection head (Head) of the network, YOLO V4 also improved the loss function. Based on the loss function IOU of YOLO V3, the CIOU-Loss function was proposed in YOLO V4:

$$CIOU = IOU - \frac{\rho^2}{c^2} - \frac{v^2}{1 - IOU + v} \tag{2}$$

$$v = \frac{4}{\pi^2}\left(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h}\right)^2 \tag{3}$$

$$LOSS_{CIOU} = 1 - CIOU \tag{4}$$

Because of the symmetry of human feet, the side-looking and down-looking characteristics of shoes are also very similar. For the images of drivers' shoes captured by the camera, loss function calculation by IOU only will lead to a serious loss of accuracy. For the loss function based on IOU, when a detection result does not intersect with the ground truth, or when multiple detection results have the same size but different positions, the calculation of the loss function will be inaccurate, which will make the weighting parameters obtained by training unreliable. However, the loss function based on CIOU can well avoid the above problems. CIOU considers the Euclidean distance between the detection result and the center of the ground truth so that different loss function values can be obtained at different positions of the detection result, making the weighting parameters obtained by training more reliable.

4 Experimental Method

4.1 Data Set Setting

Due to the wide variety of shoes, the experiment divided the shoe-wearing behaviors of drivers into five categories: high-heeled shoes, platform shoes, slippers, barefoot and flat shoes. Among them, wearing high-heeled shoes, platform shoes, slippers, and barefoot are abnormal shoe-wearing behaviors, while flat shoes are standard shoe-wearing behaviors.

The experiment collected 300 images of feet of different types at different angles for training. Among them, there are 60 images of high-heeled shoes, 60 platform shoes, 60 slippers, 60 bare feet, and 60 flat shoes.

4.2 Parameter Settings

In this paper, five deep learning models based on YOLO V4, YOLO V3, YOLO V2, Faster-RCNN and Fast-RCNN were verified through comparison with the traditional SVM model. According to the network structures of different models, corresponding parameters were set respectively to ensure the optimal detection speed and accuracy of each model. Then, the model with the best comprehensive performance was selected as the supporting framework of the method of detecting drivers' abnormal wearing of shoes proposed in this paper.

The basic parameter settings of this experiment are shown in Table 1 below:

Table 1. Basic Parameter Settings of the Experiment

Learning Model / Parameter Setting	YOLO v4	YOLO v3	YOLO v2	Faster-RCNN	Fast-RCNN	SVM
weight	608	416	416	1000	1000	1000
height	608	416	416	1000	1000	1000
batch_size	64	16	64	1	1	
learning_rate	0.00261	0.001	0.001	0.00001	0.00001	
momentum	0.949	0.9	0.9			
decay	0.0005	0.0005	0.0005	0.8	0.8	

4.3 Evaluation Standard

The method proposed in this paper divides drivers' shoe-wearing behaviors into abnormal shoe-wearing behavior including wearing high-heeled shoes, platform shoes, slippers, and barefoot, and normal shoe-wearing behavior such as wearing flat shoes, which is

a typical classification problem. Hence, the experiment adopted: recall rate (Recall), precision rate (Precision), average precision rate (AP), and Intersection-over-Union (IoU) as evaluation indicators. The calculation formula of the above evaluation indicators is as follows:

$$recall = \frac{TP}{(TP + FN)} = \frac{TP}{P} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$AP = \int_0^1 p(r)dr \approx \sum_{k=1}^{N} P(k)\Delta r(k), \tag{7}$$

where, P\p is precision rate and r is recall rate

$$IoU = \frac{DR \cap GT}{DR \cup GT} \tag{8}$$

In formula (8), DR is the detection result, and GT is the ground truth, as shown in Fig. 5:

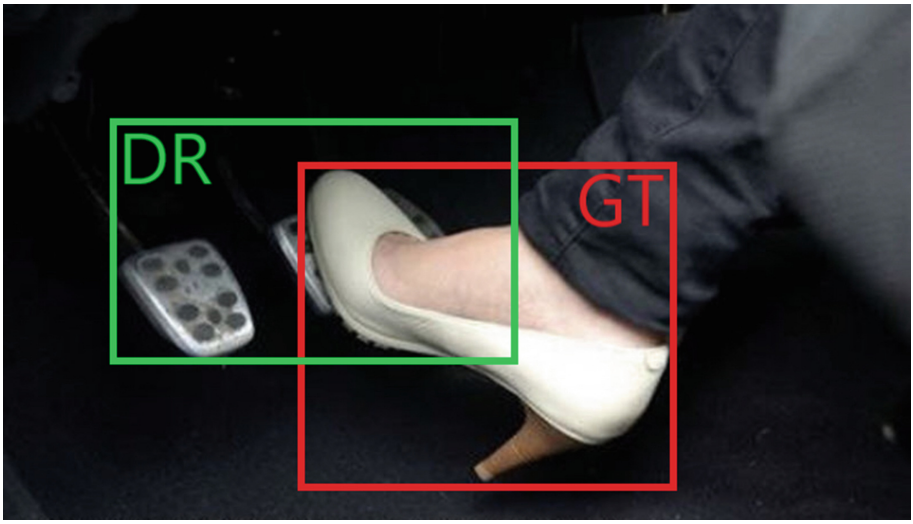

Fig. 5. DR and GT

4.4 Experiment Effect

The sample images collected in the experiment were trained by YOLO v4, YOLO v3, YOLO v2, Faster-RCNN, Fast-RCNN deep learning models and traditional SVM.

Fig. 6. Samples of Detection Results

All these models could meet the basic requirements for classification of drivers' shoe-wearing behaviors to realize detection and identification of them. Samples of detection results are shown in Fig. 6.

On this basis, the evaluation indicators of each learning model were also calculated in the experiment, as shown in Table 2 below:

Table 2. Comparison of Performance Parameters of Learning Models in the Experiment

Learning Model / Evaluation Indicators	YOLO v4	YOLO v3	YOLO v2	Faster-RCNN	Fast-RCNN	SVM
Recall	0.912	0.884	0.853	0.887	0.866	0.750
Precision	0.920	0.860	0.840	0.915	0.885	0.792
AP	0.922	0.857	0.832	0.912	0.879	0.814
IoU	0.825	0.768	0.729	0.792	0.771	0.701
FPS	75	45	48	15	12	5

As shown in Table 2 above, the four indicators of recall rates (Recall), precision rate (Precision), average precision rate (AP) and Intersection-over-Union (IoU) comprehensively reflect the accuracy of these learning models. Among the six models, YOLO V4 has the highest accuracy. YOLO retains context information when performing target recognition, which enables it to effectively avoid background misrecognition in the complex environment of the cab. The parameter FPS in Table 2 is a standard to measure target detection rate. FPS records the number of images that can be processed per second. According to the data in Table 2, the detection speed of YOLO V4 is faster than the other five models. The efficiency of YOLO V4 ensures that it will give timely feedback to the driver before driving. To sum up, YOLO V4 deep learning model is significantly better than the other five models in performance, and all the performance parameters of

the traditional SVM learning model are far lower than that of the deep learning models. Therefore, the YOLO V4 algorithm was selected as the supporting framework of the method of detecting drivers' abnormal wearing of shoes proposed in this paper.

5 Summary and Outlook

5.1 Summary

The shoe-wearing behaviors of drivers are related to the life and property safety of everyone, so an efficient and accurate detection and identification network are essential. Based on the above introduction to deep learning models and the comparison of performance among multiple learning models obtained through the experiment, it can be shown that YOLO V4 has the best comprehensive performance among all the models, proving that it is reasonable to adopt YOLO V4 algorithm as the supporting framework of the method of detecting drivers' abnormal wearing of shoes proposed in this paper. For the goal of detecting abnormal shoe-wearing behaviors of drivers, the characteristics of deep learning allow the neural network to capture more fine-grained features of each type of shoe, and the structure of YOLO V4 helps the neural network to complete feature enhancement more quickly and accurately. Based on these advantages, the neural network can classify the images of drivers' shoes captured under different rays of light and angles with high accuracy.

5.2 Future Prospects

In the experiment described in this paper, high-heeled shoes, platform shoes, slippers, barefoot and flat shoes were used for training and detection. In the future, we will continue to add new shoe types, improve the training data set, and make the neural network more universal. Moreover, in addition to the detection of drivers' shoe-wearing behaviors before driving, we will further study the detection of drivers' abnormal behaviors during driving. There are also standards of safe behaviors for drivers during driving. By training the neural network and using the deep learning model to capture and enhance the corresponding features of normal driving behaviors, theoretically, the abnormal driving behaviors of drivers can be detected and identified in the course of their driving.

References

1. Car Home. Pay attention to safety, pay attention to your feet [EB/OL] (2009). https://club.autohome.com.cn/bbs/threadowner/59ac783ba8317616/2161861-1.html
2. Yang, K., Zhang, D., Yang, L.: Vehicle driver safety belt detection based on deep learning. J. China Univ. Metrol. **3**, 326–333 (2017)
3. Xiong, Q., Lin, J., Yue, W.: A deep learning-based method for detecting driver's calling behavior. Control Inf. Technol. (6), 53–56, 62 (2019)
4. Turk, M.: Eigenfaces for recognition. J. Cogn. Neurosci. 3 (1991)
5. Chen, D., Cao, X., Wang, L., Wen, F., Sun, J.: Bayesian face revisited: a joint formulation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7574, pp. 566–579. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33712-3_41

6. Darui, S., Lenan, W., et al.: Face recognition based on nonlinear feature extraction and SVM. J. Electron. Inf. Technol. **26**(2), 307–311 (2004)
7. Omerustaoglu, F., Okan Sakar, C., Kar, G.: Distracted driver detection by combining in-vehicle and image data using deep learning. Appl. Soft Comput. **96**, 106657 (2020)
8. Lu, W., Hu, H., Wang, J., Wang, L., Deng, Y.: Tractor driver fatigue detection based on convolution neural network and facial image recognition. Trans. Chin. Soc. Agricul. Eng. **34**(7), 192–199 (2018)
9. Liu, Z., Peng, Y., Hu, W.: Driver fatigue detection based on deeply-learned facial expression representation. J. Vis. Commun. Image Represent. **71**, 102723 (2020)
10. Nojiri, N., Kong, X., Meng, L., Shimakawa, H.: Discussion on machine learning and deep learning based makeup considered eye status recognition for driver drowsiness. Procedia Comput. Sci. **147**, P264-270 (2019)
11. Wang, X., Zhang, W.: Anti-occlusion face recognition algorithm based on a deep convolutional neural network. Comput. Electr. Eng. **96**, Part A (2021)
12. Zhao, C.H., Zhang, B.L., Zhang, X.Z., Zhao, S.Q., Li, H.X.: Recognition of driving postures by combined features and random subspace ensemble of multilayer perceptron classifiers. Neural Comput. Appl. **22**(1), 175–184 (2013)
13. Xing, Y., Lv, C., Cao, D., Velenis, E.: Multi-scale driver behavior modeling based on deep spatial-temporal representation for intelligent vehicles. Transp. Res. Part C: Emerg. Technol. **130** (2021)
14. Lafuente-Arroyo, S., Martin-Martin, P., Iglesias, C., Maldonado-Bascon, S., Acevedo-Rodriguez, F.J.: RGB camera-based fallen person detection system embedded on a mobile platform. Expert Syst. Appl. **197**, 116715 (2022)
15. Naveed, H., Jafri, F., Javed, K., Babri, H.A.: Driver activity recognition by learning spatiotemporal features of pose and human object interaction
16. Wang, C., Fu, Z.: Traffic sign detection algorithm based on YOLO v2 model. Comput. Appl. **38**(S2), 276–278 (2018)
17. Shahverdy, M., Fathy, M., Berangi, R., Sabokrou, M.: Driver behavior detection and classification using deep convolutional neural networks. Expert Syst. Appl. **149**(1), 113240 (2020)
18. Xu, M., Fang, H., Lv, P., Cui, L., Zhang, S., Zhou, B.: D-STC: deep learning with spatio-temporal constraints for train drivers detection from videos. Pattern Recogn. Lett. **119**(1), 222–228 (2019)
19. Xiao, W., Liu, H., Ma, Z., Chen, W.: Attention-based deep neural network for driver behavior recognition. Future Gener. Comput. Syst. **132**, 152–161 (2022)