Calling Behavior Detection Of Port Truck Drivers Based On Deep Learning

Jing He[1], Yefu Wu[1],Jinyong Xiao[2]

1. School of Computer and Artificial Intelligence, Hubei Key Laboratory of Transportation Internet of Things,Wuhan University of Technology,Wuhan 430063,Hubei,P.R China

2.Chongqing Guoyuan Port Co.,Ltd;Chongqing 401133,China

e-mails: hej5464@126.com; wuyefu@whut.edu.cn;7405612@qq.com

Abstract——Due to the lack of a rigorous evaluation model in the traditional detection method of driver's illegal answering phone calls, it is difficult to meet the identification needs of truck drivers who answer the phone illegally in the port environment. A multi-feature fusion detection method based on deep learning is proposed. The method performs weighted fusion of the features of hand-held phone and speech to detect the calling behavior of port truck drivers. Mainly recognize faces through Retinaface training, and then extract the information of human key points through the PFLD_CPN fusion key point detection model, and use YOLOv5 target detection to identify mobile phones; determine whether there is a phone call. The experimental results show that the method can effectively detect the behavior of answering calls in real-time and effectively in the self-collected monitoring screen data of port truck drivers.

Keywords——phone call behavior detection; face key points; CPN skeleton key point location; target detection

I. INTRODUCTION

Due to the complex working environment of port truck drivers, this paper uses the improved Retinaface[1] algorithm to detect faces, uses the PFLD_CPN fusion key point detection method to detect the behavior of handheld phones, and recognizes speech behavior according to the feature information of key points of the mouth. Then uses YOLOv5 Small object detection recognizes the phone. Based on these feature information, a phone-calling behavior detection model of port truck drivers is established to determine whether the truck drivers have phone-calling behavior.

II. MODEL BUILDING

The method in this paper first collects the video data of the truck driver through the camera, and then intercepts the image frame of the video, recognizes the face through the deep learning method Retinaface face detection model, and then trains the feature point detection model of PFLD_CPN, obtains the feature location information, and calculates The degree of mouth closure, the movement of holding the phone and other characteristics, and then identify the mobile phone through YOLOv5 to determine whether there is a phone call. The flow chart 1 of this method is as follows:

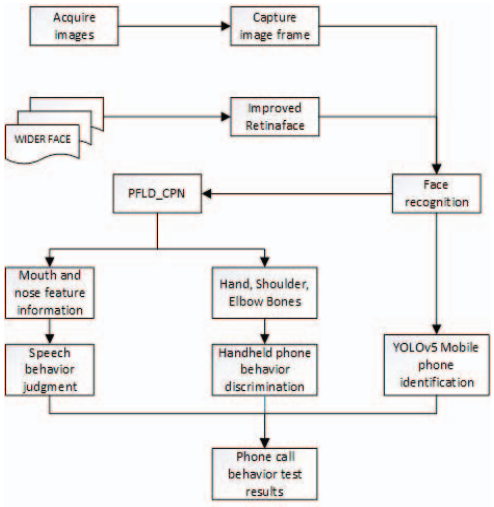

Figure 1 Flowchart of the algorithm

2473-3636/22/$31.00 ©2022 IEEE
DOI 10.1109/DCABES57229.2022.00024

III. ALGORITHM DESIGN

A. Face Recognition

RetinaFace is a robust one-stage face detection model. It learns by using the multi-task combination of additional supervision and self-supervision to perform pixel-level localization of face images of different sizes, which can adapt to face detection in different environments and has a good face recognition effect. In order to meet the real-time requirements of face detection, it is necessary to adopt the idea of In order to meet the real-time requirements of face detection, the idea of optimizing the network structure is adopted, and a lightweight network structure is adopted to reduce the amount of parameters as much as possible. The feature extraction network adopts the lightweight convolutional neural network MobileNet to replace the main feature extraction network ResNet50 of RetinaFace. The core idea of MobileNet is to use depth-wise separable convolution instead of ordinary convolution to reduce the amount of parameters of the model and improve the efficiency of real-time detection.

B. PFLD_CPN Key Point Detection

In the scene of calling behavior detection, we only need to locate the key points of the mouth and the position information of the key points of the hands and elbows. Due to the complex port environment, this paper proposes a method PFLD_CPN that combines PFLD[2] and CPN[3], removes the module for calculating Euler angles of the head in the auxiliary network in PFLD, and integrates the improved CPN human skeleton key point localization model into , locate the key points of the left and right shoulders, elbows, and hands of the human body.

In the model design, the backbone network of the PFLD model does not use VGG16, ResNet and other networks, but in order to increase the expressiveness of the model, the output features of MobileNet[4] are structurally modified, and the model expressing ability is increased by fusing features of different scales.

Aiming at detecting the positions of key points of left and right shoulders, elbows and hands, an improved CPN skeleton extraction algorithm is proposed. The algorithm adds the convolutional module-based attention mechanism to the CPN network, and adopts the CBAM module after the output of the Bottleneck block of the ResNet50 network. Through different weight assignments, the network can focus on learning useful features, and then suppress ineffective features, which strengthens the focus on feature maps of key points and reduces the focus on complex backgrounds. This can better improve the positioning accuracy of key points of the human body in complex backgrounds, and has a good detection effect for operations in the background environment of port operations.

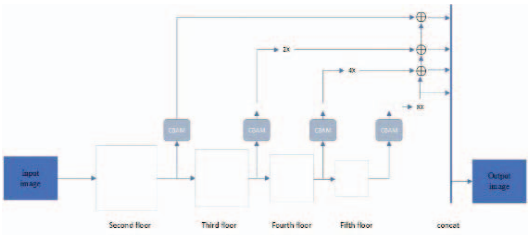

Figure 2 Introducing CBAM's CPN Network

C. Speech Behavior Recognition

For the driver's speech behavior, this paper uses the change of the aspect ratio of the mouth to identify the mouth state . According to the PFLD detection model, we can obtain the key feature points of the driver's mouth, so the mouth can be located according to the number of the feature points. and identify its status.

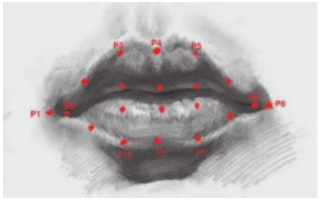

Figure 3 Schematic diagram of 10 key points in the mouth

State is judged by calculating the mouth

aspect ratio (MAR). In order to make the MAR value more accurate, the 10 feature points marked in Figure 4 are used to calculate the mouth aspect ratio:

$$MAR = \frac{Mean(Dis(P_3, P_{10}), Dis(P_4, P_9), Dis(P_5, P_8))}{Mean(Dis(P_1, P_6), Dis(P_2, P_7))} \quad (1)$$

Under normal driving conditions, the mouth is closed. When talking on the phone with others, the mouth is constantly opening and closing, and the opening and closing range is not large. When yawning when sleepy, the mouth opening is large. and lasts longer. When 0.4<MAR≤0.8, it is a normal speaking state. The state of the mouth can be determined by this feature. Since people generally continue to speak, the aspect ratio of 20 consecutive frames of images is set. If the MAR value satisfies 0.4<MAR≤0.8, the number of frames accounts for 0.7, and we judge that this is the speaking state.

D. Handheld Phone Behavior Recognition

There is a certain regularity in the angle characteristics of the hand-held phone behavior. When a person makes a call, the hand-elbow-shoulder angle is constant throughout the whole process of answering the phone, so the angle between the hand-elbow-shoulder is chosen as one of the basic characteristics of answering the phone.

The schematic diagram of the angle of hand-elbow-shoulder is shown in the figure below. According to the obtained position coordinate information of hand, elbow and shoulder, the angle between hand-elbow-shoulder is calculated, namely

$$\theta = \arccos\frac{a^2 + c^2 - b^2}{2ac}, \theta \in [0, 2\pi] \quad (2)$$

where: θ represents the angle between the hand-elbow-shoulder; a represents the distance between the elbow and the shoulder; b represents the distance between the hand and the shoulder; c represents the distance between the hand and the elbow . Since it is possible to answer the phone with both left and right hands, the positional relationship between the left (right) hand, left (right) elbow and left (right) shoulder is judged in parallel.

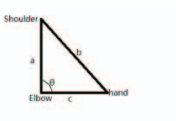

Figure 4 Shoulder-elbow- hand angle diagram

Using the extracted distance to calculate the angle information, when the angle is detected in 20 consecutive frames $\theta < 45^\circ$, it is judged that the current driver is in the state of holding the phone, and has the behavior of holding the phone.

E. YOLOv5 Mobile Phone Detection

The proportion of mobile phones in the image to be detected is relatively small, and due to the proportion of the mobile phone in the image to be detected is relatively small. Due to the complexity of the detected background, it is easy to be disturbed, and it is difficult to detect the accuracy of such small target objects. By using the YOLOv5[5] target detection algorithm, we use the position information of the key points of the hand bones to determine the detection area, and further determine whether there is a phone answering behavior by performing mobile phone detection on the target area.

IV. Experimental Results and Analysis

The experiment uses Python3.6 and OpenCV4.2 image vision library as the development environment, and conducts experiments on a computer with Intel Core i7-11800 CPU, RTX3060 GPU and 16G memory; the training sample data comes from the "state-farm-dist-racted- driver-detection" dataset.

The method in this paper detects the calling behavior of port truck drivers by detecting the driver's hand-held phone behavior, speech

behavior and various features of the method based on YOLOv5 detection of mobile phones. By calculation, the method of this paper $A_c = 96.03\%$, $R_c = 94.1\%$.

On the self-collected monitoring screen data set of port truck drivers, experiments were using HOG+SVM_RBF,CNN+ROI+YOLOv3 method and the algorithm in this paper. It can be seen that the accuracy and recall of the detection method proposed in this paper are higher than other methods (as shown in Table 1).

TABLE I. Detection results of different methods

Sample type	Average Accuracy	Recall Rate
HOG+SVM_RBF	85.63%	91.7%
CNN+ROI+YOLOv3	92.70%	93.5%
Algorithm	96.03%	94.1%

In order to test the universality and practicability of the algorithm in this paper, the method in this paper is compared with other methods of phone call behavior detection: the HOG+SVM_RBF model proposed by Bu Qingzhi[6] et al. The multi-scale fusion model detection of CNN+ROI+YOLOv3 by Xu TingTing[7]. It can be seen from Table 1 that the average recognition rate of the algorithm in this paper can reach 96.03%, which is better than the other two detection models, and the recall rate of this model is also significantly higher than that of the other two models; Figure 7 shows the 5 Average recognition rate in detections:

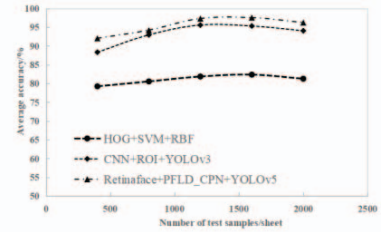

Figure 5 Recognition rates under different sample sizes

V. Concluding Remarks

This paper proposes a deep learning calling behavior detection algorithm based on Retinaface+PFLD_CPN+YOLOv5. It can judge whether the driver is making a phone call through the speech state and the behavior of the handheld phone. Compared with the previous detection methods, it has higher detection accuracy in the complex environment of the port, which can prevent false alarms to a large extent, and can be used for actual port driving scenarios. Under the driver's phone behavior, accurate and timely early warning.

VI. Acknowledgment

This research was funded by the National Natural Science Foundation of China Research and Development Program at 2021U1764262.

References

[1] Deng J , Guo J , Ververas E , et al. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[2] Guo X , Li S , Zhang J , et al. PFLD: A Practical Facial Landmark Detector[J]. 2019.

[3] Wu Q , Sun B X , Xie B , et al. A PERCLOS-Based Driver Fatigue Recognition Application for Smart Vehicle Space[C]// Third International Symposium on Information Processing. IEEE Computer Society, 2010.

[4] Howard A G , Zhu M , Chen B , et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications[J]. 2017.

[5] Qiu X , Sun X , Chen Y , et al. Pedestrian detection and counting method based on YOLOv5+DeepSORT[C]// Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. SPIE, 2021.

[6] Bu Q Z , Qiu Jun, Hu Chao . Research on the detection method of driver's scattered behavior based on HOG feature extraction and SVM [J]. Integrated Technology, 2019, 8(4): 69-75.

[7] Xu T T, Fu Junqiong, Luo Kun.Based on CNN and multi-scale fusion driver's phone call behavior detection [J]. Computer Technology and Development, 2022, 32(02): 88-93.