A Deep Learning Approach for Vehicle and Driver Detection on Highway

Peihua Lv1,2, Yang Zhang1,2, Xiaobo Lu1,2, a and Di Zhou3

1 School of Automation, Southeast University, Nanjing 210096, China
2 Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing 210096, China
3 Zhejiang Uniview Technologies Co., Ltd, Zhejiang, 310018, China

a Corresponding author: xblu2013@126.com

Abstract. The technology of the detection for vehicle and driver is a popular spot in these ten years. In particular, the driver detection is still a troubled question in the study of public security. In our paper, an algorithm based on YOLOv3 and support vector machine (SVM) is proposed for realizing the detection of vehicles on highway, as well as the detection and binary classification of people in the vehicles, so as to achieve the purpose of distinguishing drivers and passengers and form a one-to-one correspondence between vehicles and drivers. The effectiveness of the algorithm is verified under various complicated highway conditions. Compared with other advanced vehicle and driver detection technologies, the model has a good performance and is robust to road blocking, different attitudes and extreme lighting.

1. Introduction

The technology of detection has been widely spread in various kinds of fields, particularly in the field of vehicle and driver detection including the security of the public and the order of traffic. With the rapid development of modern society, various kinds of face detection techniques are being explored. While in practical application, due to the influence of different lighting, different occasions and different attitudes, the performance and effect of detection are not as good as ideal expectation.

The technology of detection has been a hot research area in deep learning in recent years. During 2004, Viola proposed a detection method named cascade, based on the Haar-Like [1] characteristics of AdaBoost [2] to execute the cascade classifier. Unfortunately, subsequent studies [3,4] have shown that it cannot maintain reliable performance in practical applications, because of the affection of people’s visual consistency. In the following years, deep CNNs are used for the technology of detection. Yang [5] offered a network with deep neural for the recognition of feature, aiming to obtain a high response rate of regions and generate candidate windows. However, the algorithm is poor in real-time. Subsequently, R-CNN [6] and fast R-CNN [7] and even faster R-CNN [8,9] were generated successively, but the network was too large and the detection speed was not high. The YOLOv3 [10] proposed by Joseph Redmon in 2018 greatly shortens detection time and improves detection efficiency while ensuring accuracy.

At the same time, driver detection has attracted extensive research interest. The research in this field can be done by Support Vector Machine (SVM). By selecting the function, the optimal classification of people can be made to distinguish passengers and driver in the vehicle.
In this paper, an improved vehicle and driver detection model based on YOLOv3 and SVM [11] is proposed, which is called IYOLO-SVM to form an adaptive detection model. The model is trained, based on our own database which is composed by pictures of traffic vehicle and driver. These pictures are provided by the Jiangsu Provincial Transportation Department. The result of the IYOLO-SVM model is verified on another database for testing. The finally experimental results show that compared with other advanced methods, IYOLO-SVM model can detect both drivers and the vehicle, and classify people in the vehicle, which has a higher detection rate and a lower error detection rate.

2. Basic Theory
In this article, we have improved and optimized the network structure of YOLOv3 and cascade SVM at the end of the network, which greatly improve the original network. In the next part of the article, we will briefly introduce the structure of the network for detection and the definition of the SVM classifier.

2.1 YOLOv3
The core idea of YOLOv3 is using the picture as a network input, return directly to the bounding box and its dependent categories at the output layer [12]. The full stages of YOLOv3 which is composed of four parts are shown in the following paper.

2.1.1 Bounding Box Prediction
The anchor boxes of YOLOv3 are made by clustering. The values of four coordinates [13] for each of the bounding box prediction (tx, ty, tw, th) in predicting cell (a picture into S * S grid cells) based on the left top corner of the picture offset (cx, cy), according to the bounding box of pw, ph width and height, the bounding box can be predicted as follows:

\[
\begin{align*}
    b_x &= \sigma(tx) + cx \\
    b_y &= \sigma(ty) + cy \\
    bw &= pw \times e^m \\
    bh &= ph \times e^n \\
    m &= tw, n = th
\end{align*}
\]

The sum of the error loss of square is used to predict the coordinate value, so the error can be calculated rapidly.

YOLOv3 predicts the score of an object for each bounding box by logistic regression. If the prediction of the bounding box and the real border overlap better than that of the other all forecasts, then the value equals 1. If the overlap does not get the value of a threshold (setting 0.5), the prediction of bounding box will be neglected, and is displayed as no loss.

2.1.2 Class Prediction
To classify different kinds of objections, independent logistic classifiers are used instead of a SoftMax. When training, binary cross-entropy loss is used for the class predictions.

2.1.3 Predictions Across Scales
YOLOv3 predicts different boxes in three different scales. YOLOv3 uses FPN (feature pyramid network) to extract feature from scales, and finally predicts a 3-d tensor, containing the bounding box information, object information, and class information.

2.1.4 Feature Extractor
YOLOv3 uses a complex network for performing extraction, which has 53 convolutional layers, called Darknert-53.

2.2 Support Vector Machine
Support Vector Machine (SVM) is a category of generalized linear classifier that classifies data according to supervised learning. The decision boundary is the maximum margin hyperplane for solving
learning samples. In order to extend the functionality of our model, we use SVM to classify the two types of objects to determine whether the target is the driver or the passenger, to achieve multiple classification judgments. Experiments show that the connection between SVM and YOLOv3 shows strong ability for detection and classification in different types of environments.

3. Our Improved Method

3.1 The Whole Circuit of IYOLO-SVM

The figure below shows the circuit of the algorithm of IYOLO-SVM. At first, the IYOLO-Net obtains a large number of bounding boxes of cars and people from the input image. Then, the precise regions of vehicle and people are confirmed. After that, the classification model is used to differ whether it is a driver or a passenger and then give labels. Finally, the picture with bounding boxes and labels is expert.

Figure 1. The Whole Circuit of IYOLO-SVM algorithm.

3.2 The Structure of Our New Network

In our IYOLO-SVM algorithm, we have made an obviously improvement on the structure of primary network, which becomes smaller and more efficient (Table.1). The required size for input image is 416*416. There are totally four maxpool layers and twenty-two convolutional layers. In the network [14], the role of the routing layer is to introduce finer-grained granularity features from earlier locations in the network. The reorg layer matches these features to the feature map of the next layer. The size of the end feature map is 13 * 13, the size of the previous feature map is 26 * 26 * 512. The reorganized layer maps the 26 * 26 * 512 feature map to the 13 * 13 * 2048 feature map so that it can be mapped to 13 * 13 resolution function chart. Through this method, high resolution features and low-resolution features are linked together, which can increase the recognition accuracy of small objects such as our people in the vehicle.
### Table 1. The Structure of Our Network.

| Number | Layer   | Filter | Measurement | Out  |
|--------|---------|--------|-------------|------|
| 0      | CONV.   | 32     | 3^2/1       | 416^2|
| 1      | Maxpool |        | 2^2/2       | 208^2|
| 2      | CONV.   | 64     | 3^2/1       | 208^2|
| 3      | Maxpool |        | 2^2/2       | 104^2|
| 4      | CONV.   | 128    | 3^2/1       | 104^2|
| 5      | CONV.   | 64     | 1^2/1       | 104^2|
| 6      | CONV.   | 128    | 3^2/1       | 104^2|
| 7      | Maxpool |        | 2^2/2       | 52^2 |
| 8      | CONV.   | 256    | 3^2/1       | 52^2 |
| 9      | CONV.   | 128    | 1^2/1       | 52^2 |
| 10     | CONV.   | 256    | 3^2/1       | 52^2 |
| 11     | Maxpool |        | 2^2/2       | 26^2 |
| 12     | CONV.   | 512    | 3^2/1       | 26^2 |
| 13     | CONV.   | 256    | 1^2/1       | 26^2 |
| 14     | CONV.   | 512    | 3^2/1       | 26^2 |
| 15     | CONV.   | 256    | 1^2/1       | 26^2 |
| 16     | CONV.   | 512    | 3^2/1       | 26^2 |
| 17     | Maxpool |        | 2^2/2       | 13^2 |
| 18     | CONV.   | 1024   | 3^2/1       | 13^2 |
| 19     | CONV.   | 512    | 1^2/1       | 13^2 |
| 20     | CONV.   | 1024   | 3^2/1       | 13^2 |
| 21     | CONV.   | 512    | 1^2/1       | 13^2 |
| 22     | CONV.   | 1024   | 3^2/1       | 13^2 |
| 23     | CONV.   | 1024   | 3^2/1       | 13^2 |
| 24     | CONV.   | 1024   | 3^2/1       | 13^2 |
| 25     | Route   |        |             |      |
| 26     | Reorg   |        |             |      |
| 27     | Route   |        |             |      |
| 28     | CONV.   | 1024   | 3^2/1       | 13^2 |
| 29     | CONV.   | 40     | 1^2/1       | 13^2 |
| 30     | Detection |       |             |      |

### 3.3 Manual Hard Sample Mining

The traditional method for handling difficult samples is to manually screen out the difficult samples which cannot be classified after self-inspection through the training network. This traditional method [15] is slow and inefficient. This method adopts the method of online difficult sample back propagation. In each mini-batch, the calculated loss is sorted from forward propagation of total samples, then only top seventy percent of the loss is taken as the difficult sample. Then, in the back propagation, only the difficult samples are calculated, and the simple samples are ignored. The online difficult sample back propagation greatly reduces the artificial labour force and improves the training efficiency.

We use the database built by our own to train the IYOLO-SVM model. Similarly, we also use our own testing set for verification. What we need to pay attention to is that, even if we have added online hard sample mining to the network. However, in actual situations, especially in complex highway situations, because of complex lighting changes, attitude changes, object occlusion and so on [16], it is significant to randomly add difficult samples, for the aim to increase the accuracy detection rate and drop the false detection rate in the final experiment. The final results show that this method makes the model improve performance apparently, which is showing in the next section.
4. The Results After Experiments

This following chapter, we compose our own image database through photos provided from the Jiangsu Provincial Transportation Department. It contains about 1,500 photos of the driver's travel in different situations containing their faces and cars. We randomly selected 1200 pictures for training our IYOLO-SVM model and 300 pictures for testing the performance of IYOLO-SVM model. GPU GTX1050ti is applied to train the IYOLO-SVM networks. The convergence graph of loss and results of our model on detection and classification are shown below.

4.1 The Procedure of Training Loss

Figure 2 shows the procedure of training IYOLO-SVM model. The picture indicates that the abscissa stands for the number of times of iterations during training. The ordinate stands for the loss of IYOLO-SVM model. When there are objects in the grid, confidence loss of bounding boxes calculates the weight of the contribution to the total loss for five; When there is no object in the grid, confidence loss of bounding boxes calculates the contribution weight to the total loss for one; The weight of the contribution of category loss to the total loss is calculated for one; The weight of the contribution of bounding boxes’ coordinates prediction loss to the total loss is calculated for one.

![Figure 2. The procedure of training IYOLO-SVM model.](image)

4.2 The Performances of Detection and Classification

The two groups of pictures below reveal the final testing results of the IYOLO-SVM model in different kinds of complex environments. According to the final testing pictures of the IYOLO-SVM model in day and night, the model can be found to be robust under illumination changes. Table 2 is a comparison table comparing the testing results of IYOLO-SVM, YOLOv3 and Cascade CNN. From the experimental testing results, it is obvious to find that the improved method raised by us owns a higher accuracy detection rate while maintaining a lower error detection rate when compared with other methods. In addition, the model also has a good classification effect by using support vector machine. The above phenomena all show that the IYOLO-SVM model greatly improve the accuracy of vehicle and driver detection on complex highway.
Figure 3. The detection and classification consequences in night environments under the IYOLO-SVM model.
5. Conclusions

This paper raises a model of vehicle and driver detection based on improved YOLOv3 and SVM. As there might be more than one person in the vehicle, which driver detection through convolutional network in YOLOv3 cannot be realized, it is meaningful to use support vector machine for the judgement of driver and passenger. At the same time, the IYOLO-SVM model also provides a high-accuracy vehicle and driver detection technology in the complex environment on the highway. The final experimental results reveal that our new model raised in this paper is an efficient model with higher detection rate, lower error detection rate, fast detection speed, short training time and strong robustness to complex environments and different illumination on the highway.

However, our model has a bad performance when there is too much object occlusion and the driver's pixel is too small. Therefore, in the future, we need to solve and improve these problems for further study.

Acknowledgments

I would like to thank the National Natural Science Foundation of China (No.61871123), the Jiangsu Provincial Key Research and Development Program (No.BE2016739) and the Jiangsu Provincial University's Priority Academic Project Development Funding Project for our research support. Thanks to the lab's brothers and sisters and teachers for their help in my experiment.
References
[1] Mita, T., T. Kaneko, and O. Hori. “Joint Haar-like features for face detection.” Tenth IEEE International Conference on Computer Vision (2005)
[2] Zhu, Ji, et al. “Multi-class AdaBoost.” Statistics & Its Interface 2.3 pp.349-360 (2006)
[3] Q. Zhu, M. C. Yeh, K. T. Cheng, and S. Avidan, “Fast human detection using a cascade of histograms of oriented gradients,” in IEEE Computer Conference on Computer Vision and Pattern Recognition, pp.1491-1498 (2006)
[4] M. T. Pham, Y. Gao, V. D. D. Hoang, and T. J. Cham, “Fast polygonal integration and its application in extending haar-like features to improve object detection,” in IEEE Conference on Computer Vision and Pattern Recognition, pp. 942-949 (2010)
[5] S. Yang, P. Luo, C. C. Loy, and X. Tang, “From facial parts responses to face detection: A deep learning approach,” in IEEE International Conference on Computer Vision, pp. 3676-3684 (2015)
[6] Girshick, Ross. “Fast R-CNN,” Computer Science, (2015)
[7] Wang, Xiaolong, A. Shrivastava, and A. Gupta. “A-Fast-RCNN: Hard Positive Generation via Adversary for Object Detection,” IEEE Conference on Computer Vision and Pattern Recognition IEEE, pp.3039-3048 (2017)
[8] Ren, Shaoqing, et al. “Faster R-CNN: towards real-time object detection with region proposal networks,” International Conference on Neural Information Processing Systems MIT Press, pp.91-99 (2015)
[9] Fan, Quanfu, L. Brown, and J. Smith. “A closer look at Faster R-CNN for vehicle detection,” Intelligent Vehicles Symposium IEEE, pp.124-129. (2016)
[10] Redmon, Joseph, and A. Farhadi. “YOLOv3: An Incremental Improvement,” (2018)
[11] Suykens J A K, Vandewalle J. “Least Squares Support Vector Machine Classifiers,” Neural Processing Letters, 9(3):293-300 (1999)
[12] Liu, Xueliang, et al. “Correction to: On fusing the latent deep CNN feature for image classification.” World Wide Web (2018)
[13] Ranjan, Rajeev, V. M. Patel, and R. Chellappa. “HyperFace: A Deep Multi-task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition.” IEEE Transactions on Pattern Analysis & Machine Intelligence pp.99:1-1(2018)
[14] Foroozandeh Shahraki, Farideh. “Cyclist Detection, Tracking, and Trajectory Analysis in Urban Traffic Video Data.”, University of Nevada, Las Vegas, (2018)
[15] Nguyen, D. T., et al. “Combining Deep and Handcrafted Image Features for Presentation Attack Detection in Face Recognition Systems Using Visible-Light Camera Sensors.” Sensors 18.3:699. (2018)
[16] Yang, Shuo, et al. “Face Detection through Scale-Friendly Deep Convolutional Networks.” (2017)
[17] Yang Zhang, Peihua Lv, Xiaobo Lu. “A Deep Learning Approach for Face Detection and Location on Highway”, IOP Conference Series: Materials Science and Engineering, (2018)