Vehicle and Driver Detection on Highway Based on Cascade R-CNN

Yifei Geng1, 2 and Xiaobo Lu1, 2, 3
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 Corresponding author: xblu2013@126.com

Abstract. Vehicle and driver detection in the highway scene has been a research hotspot in the field of object detection in recent years, and it is still a challenging problem in the research of traffic order and road safety. In this paper, we propose a novel end-to-end vehicle and driver detection method named VDDNet which is based on Cascade R-CNN and SENet. By introducing FPN structure and SENet attention mechanism in the backbone, the ability of the model to learn effective features is enhanced. It can improve the accuracy of detection in difficult scenes such as weak light, partial occlusion, and low picture resolution. The test results based on the database of highway traffic vehicle and drivers constructed by the Jiangsu Provincial Public Security Department. It shows that the detection method has the AP rate of 91.3% and the Recall rate of 92.4%, which demonstrates the effectiveness of the proposed method in complex highway environments.

1. Introduction

With the increasing concern of society for road safety and traffic order, intelligent transportation has become a research hotspot in the field of artificial intelligence. Intelligent transportation can not only effectively maintain traffic order, but also reduce environmental pollution, ensure traffic safety, and improve transportation efficiency. The intelligent transportation system includes three parts: intelligent vehicles, intelligent highway systems, and intelligent drivers. Therefore, research on vehicle and driver detection in highway scenarios is of great value for the development of intelligent transportation systems and intelligent vehicles.

In terms of vehicle detection technology, Alonso presented a multi-dimensional classification method to realize the detection of vehicles on traffic roads [1]. Chang and Cho used a vehicle detection method based on an online boost, which solved the problem of vehicle detection in different scenarios [2]. With the rise of deep learning technology, more and more researchers are trying to bring it into their own research fields. Zhang employed a vehicle detection method based on a deep convolutional neural network, which can accurately detect vehicles in natural scenes [3]. Zhang QH presented an improved vehicle detection method based on an improved fast regional convolutional neural network (R-CNN), which improved the detection accuracy and reduced the detection time [4]. In terms of human detection technology, Ouyang et al. used a joint deep learning method to learn feature extraction, deformation, and occlusion processing models and classifiers in a unified convolution framework [5]; Mao et al. used the Hyper-Learner network to jointly learn the detection objects and specific additional features [6]; Zhao et al. proposed a method of combining RPN with a fast advancement tree in which the detection accuracy was slightly improved [7].
However, in actual scene applications, due to the instability of illumination intensity, changes in driver attitude, and the limitation of camera resolution, the detection performance of current models is significantly affected, leading to bad detection results. Therefore, this paper proposes an improved vehicle and driver detection model based on Cascade R-CNN [8], Resnet-101 [9], and SENet [10], called VDDNet. The model was trained on the driver database of highway traffic vehicles, which is constructed by the Jiangsu Provincial Public Security Department, and the performance of the test data set was verified. The experimental results show that compared with Yolov3 [11] and Cascade R-CNN, VDDNet model is more robust to challenging scenarios, and it is more accurate for object classification and regression.

2. The Proposed Method
In the experiment, we found that existing networks such as Yolov3 or Cascade R-CNN could not detect accurately in difficult scenarios. Because the network cannot fully learn the differential characteristic between the object and the background, the detected object and the background cannot be distinguished precisely, resulting in a lower accuracy rate. In order to improve the robustness of the model in difficult scenarios, we realize the fusion of Cascade RCNN and SENet and propose the VDDNet. In this section, the basic theories of Cascade R-CNN and SENet are introduced. Also, our proposed VDDNet and training tricks will be demonstrated.

2.1. Cascade R-CNN Network Structure
Cascade R-CNN is a multi-stage extension of the popular two-stage R-CNN object detection framework. The purpose is to obtain high-quality object detection, which can forcefully reject close false positives. It consists of a series of end-to-end trained detectors whose IoU thresholds are continually increasing, making them more selective for near false positives. The output of a previous stage detector is forwarded to a later stage detector, and the detection results will be improved stage by stage.

2.2. Feature Pyramid Network (FPN)
In this experiment, to deal with small object detection and obtain rich context information, an FPN structure was added to Cascade R-CNN. FPN uses the inherent multi-scale, pyramid hierarchies of deep convolutional networks to construct feature pyramids with little extra cost and thus proposes a top-down architecture with horizontal connections for building advanced semantic features on all scales Mapping. It can effectively empower conventional CNN models so that it can generate more expressive feature maps for use in the next stage of computer vision tasks. In essence, it is a method to strengthen the CNN feature expression of the backbone network.

2.3. SENet Network Structure
The core of SENet lies in the explicit modeling of the interdependence between feature channels. It adopts a learnable feature re-scaling strategy to increase the weight of effective feature vectors and reduce the weight of invalid feature vectors. SENet mainly uses Squeeze and Excitation as two key operations. The operation process is shown in figure 1.

Firstly, SENet performs global average pooling on the input feature map to obtain a feature map with a size of C×1×1 (C is the number of feature map channels). Then, the two full connection layers are used to reduce and raise the dimension of the feature map. What’s more, it is activated by the sigmoid function to obtain a weight of C×1×1, which is multiplied with the original input feature map at the corresponding position. The result of the final multiplication is output. SENet introduces original information into the deep layer to suppress information degradation. Global average pooling can not only expand the receptive field, but also integrate multi-angles of shallow and deep information. The combined output contains multiple levels of information, which enhances the expressiveness of the feature map.
2.4. The Whole Procedure of VDDNet

Figure 2 performs the procedure chart of the VDDNet algorithm. First, the backbone of SE-Resnet101 is used to perform feature extraction on vehicles and people in the input image. Second, to handle small objects detection and obtain rich contextual information, the feature map is imported into the FPN network, and the feature map is fused to obtain feature maps of different scales. Then, feature maps of different scales are inputted into the RPN network to get the rough rectangle proposals for the object to be inspected. Finally, these proposals are imported to R-CNN networks with different thresholds for more accurate object BBox regression and classification. And finally, the result is more precise detection and classification of vehicles and people in awkward scenes.

2.5. Online Hard Examples Mining (OHEM)

As the result of the complexity of the highway scenario environment, it is necessary to introduce the OHEM [12] mechanism, which can improve training efficiency and enhance model performance.

Thus, in this paper, we introduce the OHEM mechanism during the training phase. In each SGD iteration, only a small number of images enter the network for learning. For thousands of proposals contained in each image, only hard samples will be selected. The samples are ranked by calculating the loss value of each R-CNN. Only the first 70% of the loss value is taken as the difficult sample, and the loss value of the non-hard sample is set to zero. In this paper, the con5_x (ResNet-101) residual structure is applied. It implements downsampling to 1 * 1 pixels and then upsampling. Therefore, the network can get not only more detailed information but also contextual information. In the OHEM-R-CNN network, two fully connected layers, and ROI classification layer and an ROI position regression layer, are used to calculate the loss value.
2.6. Warmup Strategy
In the training process, we usually use the standard step-down learning rate. The initial learning rate is 3.5e-4. If we train a total of 1,000 epochs, the learning rate decreases at the 400 and 700 epochs. Because the initial gradient of the network is large, this strategy of initializing the network with a large learning rate may cause the network to oscillate to a suboptimal space. The Warmup strategy [13] is equivalent to initializing the network with a gradually increasing learning rate in the initial stage and progressively initializing it to better search space for the next training. This paper uses the most straightforward linear strategy; that is, the first 50 epoch learnings gradually increase from 0 to the initial learning rate.

3. Experimental Results
In this paper, this method is implemented by the PyTorch and the GPU is NVIDIA GeForce RTX 2080Ti. The performance of our method on detection and classification will be shown later.

3.1. Dataset Construction
In this section, we use our own traffic driver face database, which is constructed by the Public Security Department of Jiangsu Province, as a dataset, containing approximately 1500 images in different traffic conditions. In order to further enhance the diversity of the data set, we use the original image to enlarge (image width and height to 3/2), reduce (image width to 1/3, height to 1/2, and ensure that the image size is a multiple of 32), brightness enhancement, brightness reduction, flip (90 ° and 180 °) and increase noise to enhance the data, and finally obtain 9,000 pictures. We randomly select 7,200 images to train the VDDNet model, and the remaining 1800 images to verify the performance of the model.

3.2. Vision Comparison
Figure 3, figure 4 and figure 5 show the detection results of Cascade R-CNN and VDDNet models under difficult environment (for example poor light, partial occlusion, sample blur, etc.) As can be seen from the figure, compared with Cascade R-CNN, the model proposed in this paper can improve the confidence of the object to be detected and optimize the positioning quality of the bounding box regression (figure 3), reduce false detection and improve accuracy (figure 4), and reduce leakage inspection to improve recall (figure 5). We can get the conclusion that the proposed model in this paper can achieve an excellent performance in a complex environment on the highway.

![Figure 3. Detection results of Cascade RCNN (left) and VDDNet (right)](image-url)
Figure 4. Detection results of Cascade RCNN (left) and VDDNet (right)

Figure 5. Detection results of Cascade RCNN (left) and VDDNet (right)

3.3. The Performances of Detection and Classification

Table 1 is a comparison table of the evaluation indicators, showing the detection performance of VDDNet, Cascade R-CNN and YOLOv3 in this test set. Compared with Cascade R-CNN, VDDNet has significantly improved on AP: 0.3% on car, 2.1% on the driver, and 4.5% on person AP. And at the same time, the recall rate has also strengthened: 0.1% on car, 1.2% on the driver, and 5.5% on person Recall. The experiment result shows that the FPN structure and SENet attention mechanism enables the model to learn the characteristics of the object comprehensively and strengthen the effective features, thus reducing the false positives. As a result, the AP is improved, false alarms are reduced, and the recall rate is improved. It is worth emphasizing that in difficult scenarios (e.g., dark light (figure 4,5), occlusion problems (figure 4), and low image resolution (figure 3)), the improvement of these indicators is more significant. It can be seen that the improved model has better accuracy and robustness in challenging scenarios. It sufficiently proves the excellent performance of the VDDNet algorithm in improving the detection accuracy of highway vehicles and drivers.

Table 1. Performances of detection between VDDNet and other comparison techniques

| Method         | AP (%) | Recall (%) | mAP |
|----------------|--------|------------|-----|
|                | car    | driver     | person | car | driver | person |       |
| Yolov3         | 84.3   | 74.5       | 56.3   | 85.6 | 78.9    | 66.1    | 71.7   |
| Cascade RCNN   | 91.0   | 76.8       | 66.7   | 92.3 | 81.4    | 73.9    | 78.2   |
| VDDNet         | 91.3   | 78.9       | 71.2   | 92.4 | 82.6    | 79.4    | 80.5   |
4. Conclusion
Using Resnet-101 as the backbone of the network, by introducing the FPN structure and the SENet attention mechanism into the backbone, the model can thoroughly learn the features of the object to be detected and enhance the weight of useful features. Therefore, the features under challenging scenes are more prominent, which effectively improves the accuracy of the detection in the next stage, and achieves accurate detection of vehicles and drivers in complex highway environments. The experimental results show that the VDDNet proposed in this paper has good robustness to occlusion, attitude change, and extreme lighting of highways. However, our work is still not accurate enough to detect passengers in complex lighting environments and shadows and is slower in detection speed. Therefore, in future work, model pruning and model compression are a direction worth improving.

5. Acknowledgments
This work was supported by the National Natural Science Foundation of China (No.61871123), Key Research and Development Program in Jiangsu Province (No.BE2016739) and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

6. References
[1] Alonso D, Salgado L and Nieto M 2007 Robust vehicle detection through multidimensional classification for on board video based systems. IEEE Int. Conf. on Image Processing (San Antonio) vols 1-7 pp 2017-20.
[2] Chang WC and Cho CW 2010 Online boosting for vehicle detection. IEEE Trans. Syst. Man Cybernet Part B-Cybernetics 40 pp 892-902.
[3] Zhang FY, Xu X and Qiao Y 2015 Deep classification of vehicle makers and models: the effectiveness of pre-training and data enhancement. IEEE Int. Conf. on Robotics and Biomimetics (Zhuhai, China) pp 231-36.
[4] Zhang QH, Wan CX and Han WL 2019 A modified faster region-based convolutional neural network approach for improved vehicle detection performance. Multimedia Tools and Applications 78 pp 29431-46.
[5] Ouyang WL and Wang XG 2013 Joint deep learning for pedestrian detection. IEEE Int. Conf. on Computer Vision (Sydney) pp 2056–63.
[6] Mao JY, Xiao TT, Jiang YN and Cao ZM 2017 What can help pedestrian detection. 30th IEEE Int. Conf. on Computer Vision and pattern recognition (Honolulu, HI) pp 6034–43.
[7] Zhao JX, Li J and Ma YD 2018 RPN plus fast boosted tree: Combining deep neural network with traditional classifier for pedestrian detection. 4th Int. Conf. on Computer and Technology Applications (Istanbul) pp 141-50.
[8] Cai ZW and Vasconcelos N 2018 Cascade R-CNN: delving into high-quality object detection. 31st IEEE Conf. on Computer Vision and pattern recognition (Salt Lake City) pp 6154-6162.
[9] He KM, Zhang XY, Ren SQ and Sun J 2016 Deep residual learning for image recognition. IEEE Conf.on Computer Vision and pattern recognition (Seattle, WA) pp 770-78.
[10] Hu J, Shen L and Sun G 2018 Squeeze-and-excitation networks. 31st IEEE Conf. on Computer Vision and pattern recognition (Salt Lake City) pp 7132-41.
[11] Redmon J and Farhadi A 2018 YOLOv3: An Incremental Improvement. arXiv: 1804.02767 [cs.CV].
[12] Shrivastava A, Gupta A and Girshick R 2016 Training region-based object detectors with online hard example mining. IEEE Conf. on Computer Vision and pattern recognition (Seattle, WA) pp 761-69.
[13] Goyal P, Dollar P, Girshick R, Noordhuis P, Wesolowski L, Kyrola A, Tulloch A, Jia YQ and He KM 2017 Accurate, large minibatch SGD: Training imagenet in 1 hour. arXiv: 1706.02677 [cs.CV].