Driver Detection Based on Deep Learning

Mingqi Lu\textsuperscript{1,2}, Yaocong Hu\textsuperscript{1,2} and Xiaobo Lu\textsuperscript{1,2,*}

\textsuperscript{1}School of Automation, Southeast University, Nanjing 210096, China
\textsuperscript{2}Key Laboratory of Measurement and Control of CSE, Ministry of Education, Southeast University, Nanjing 210096, China
\textsuperscript{*}E-mail: xblu2013@126.com

Abstract. Driver detection is an essential part of monitoring and identifying drivers’ bad driving behavior and preventing traffic accidents. The complicated background of the cockpit, the different light and dark light and the changeable driver pose bring great difficulties to the driver detection. In this paper, a driver detection algorithm based on improved Faster R-CNN is proposed based on RGB images captured by vehicle-mounted cameras. The residual structure is introduced into the ZF network to design the ZF-Net network, in order to improve the accuracy and maintain the real-time performance. LRN (Local Response Normalization) is replaced by BN (Batch Normalization), which simplifies parameter adjustment and accelerates network convergence. Experiments on a self-built driver image database demonstrate the effectiveness of the driver detection method.

1. Introduction

The traffic accident of 80%-90% is caused by the negligence of the driver and the inobservance the traffic rules, so it is an urgent need to detect and locate the driver accurately and quickly from the vehicle-monitoring image for identifying the bad driving behavior and preventing traffic accidents. At present, there are few researches on driver detection.

With the outstanding performance of deep learning in the field of computer vision, the convolution neural network CNN (convolutional neural networks) shows superior performance on the target detection task [4-5]. The framework of CNN detection mainly includes region based method and regression based method [5]. Among them, the region based method represented by R-CNN [1] and Faster R-CNN [3] have higher detection accuracy than that based on regression method. Faster R-CNN integrates RPN (Region Proposal Networks) and Fast R-CNN [2] to achieve high accuracy real-time detection, which has received great attention. As the number of CNN network structures used by Faster R-CNN is from shallow to deep, including ZF [6], VGG [7], GoogleNet [8], and ResNet [9], even deeper networks may bring higher precision, but the detection speed will be reduced. Therefore, for specific problems, it is one of the main research directions of CNN to study the appropriate basic network structure and training methods to ensure high accuracy and ensure real-time performance [9-10].

In order to solve the problem of driver detection under the background of complex cockpit, RGB image taken by the vehicle camera are used. By introducing the residual structure and the thought of Batch Normalization [11], the driver detection method based on the improved Faster R-CNN is proposed, that is, the training network of the over to end to realize the driver's detection. Location.

2. Driver Detection Model

Aiming at the driver detection problem, Faster R-CNN algorithm is improved in this paper. The flow chart of Faster R-CNN algorithm is shown in figure 1.
Convolution network can effectively extract sources, and the network structure is shown in Figure 1. The number of channels in the output characteristic graph is the same as the size. The residual structure is introduced into the ZF network and the ZF-Res network is designed in order to obtain higher recognition accuracy and keep the real-time performance of recognition. (2) LRN (Local Response Normalization) is replaced by BN (Batch Normalization), which simplifies parameter adjustment and accelerates network convergence.

2.1. ZF-Res Network
Stacking the coiling layer on the CNN shallow network usually improves the accuracy of image recognition [7]. However, the precision gain brought by the network depth is limited by the network structure, the limited data set and the backpropagation training algorithm, and even faster with the network deepening. The application of residual structure in deep CNN network can effectively alleviate the problem of gradient disappearance and training degradation in the deep network training without increasing the cost of computing, thus improving the convergence performance of the network and the recognition precision [9]. In order to improve the detection precision and keep the real-time of the driver, a new ZF-Res (ZF with residual learning frameworks) network is designed on the basis of the ZF network with a strong real-time network, and the residual structure is introduced. Residual function is Y=F(X) +X. In the formula, X is the convolution characteristic of input residual structure. F (x) is the output of the convolution feature that is skipped by the shortcut path, which is the output of the residual structure. X and F (X) need to satisfy the equal size of the output feature dimension, that is, the number of channels in the output characteristic graph is the same as the size. The design of ZF-Res network structure is shown in Figure 3.
the number of channels of Conv2 and Conv4 in the ZF network is 256 and 384 respectively. In order to satisfy the design conditions of X and F (X) in the residual structure and increase the coiling layer as little as possible, the convolution layer of convolution kernel is 3 * 3 and the output channel is 256 of the convolution layer between Conv2 and Conv3, in Conv4 and Conv5. The convolution layer Conv4_2 with convolution kernel size of 3 x 3 and output channel 384 is increased. Compared with other convolution kernels, the 3 * 3 convolution kernel has many advantages, such as less parameters and faster speed. It has been widely used in major mainstream CNN networks. The step length of the newly added coiling layer is set to 1, and zero-padding of the 1 pixel boundary is carried out to keep the size of the output feature map after the new layer convolution is kept.

2) The number of channels in the Conv5 layer is expanded from 256 to 512, so that richer feature information can be passed back.

3) The two convolution layers of Conv2_2 and Conv2_3 and the two convolution layers of Conv4 and Conv4_2 form the residual structure by the shortcut connection, which makes the two layers learning the residual between the input and output, and forms the ZF-Res network.

**2.2. Batch Normalization**

LRN (Local Response Normalization) is replaced by BN (Batch Normalization) to achieve data normalization.

\[
\hat{x}^{(k)} = \frac{x^{(k)} - E(x)}{\sqrt{\text{Var}(x^{(k)})}}
\]

Among them, \(E(x)\) and \(\sqrt{\text{Var}(x^{(k)})}\) are the mean and variance of mini-batch respectively. However, the normalized data will affect the feature learning of the network. Therefore, the BN algorithm proposes the idea of "transformation reconfiguration", that is, the introduction of parameters \(\gamma\) and \(\beta\), which are updated iteratively in the network training.

\[
y^{(k)} = \gamma^{(k)} x^{(k)} + \beta^{(k)}
\]

It is not difficult to recover the characteristic distribution required by the original network when the satisfaction formula (3) and (4) are satisfied.

\[
\gamma^{(k)} = \sqrt{\text{Var}(x^{(k)})}
\]

\[
\beta^{(k)} = E(x)
\]

**2.3. Modification of the Parameters of Faster R-CNN**

1) The purpose of this study is to detect and locate drivers in the cockpit, so the classified categories of self-built data sets are set to 2 categories, the drivers and the background.
2) In training, the open trained ImageNet classification model is used to initialize the network layer shared by RPN and Fast R-CNN. In this paper, a small batch random gradient descent method is used to train Faster R-CNN in the end to end joint mode. The size of the mini-batch is set to 64 and the impulse is 0.9. The attenuation coefficient of the weight is $5^{-4}$, and the maximum iteration number is $8 \times 10^5$, of which the first $5 \times 10^5$ learning rate is $10^{-4}$, and the $3 \times 10^5$ learning rate is $10^{-5}$. In the initialization of the model, the shared convolution layer is initialized randomly with Xavier $^{[1]}$, while the RPN's convolution layer and the full connection layer in the network are initialized randomly with a Gauss distribution of 0 mean and a standard deviation of 0.01. In each sliding window location, RPN network takes 9 anchor points of 3 kinds of area scale and 3 aspect ratio to locate drivers.

3. Experiment

3.1. Experimental Data

In this paper, the self-built sample library simulates the driving status of drivers in the cockpit, which includes 62 people. Each person completes normal driving, smoking, calling, playing mobile phone, talking with the neighbour and disconnecting the steering wheel with six kinds of behavior. The self-built sample library also considers the degree of illumination in the cockpit, the change of perspective and so on. A total of 16750 samples have been built.

3.2. Experimental Environment

| Table 1. Experimental environment of driver detection |
|-----------------------------------------------------|
| CPU | Intel Xeon E5-2643,3.4GHz |
| GPU | Nvidia GTX 1080,8GB GDDR5 |
| Operating system | Ubuntu 16.04 |
| Deep learning software library | PyTorch |
| Programming language | Python2.7 |
| Development platform | Eclipse Luna Service Release 2 (4.4.2) |

3.3. Experimental Results and Analysis

On the self-built image dataset, the ZF, VGG-16 and ZF-Res models are used as the feature extraction network of Faster R-CNN for driver detection. The following are the experimental results:

![Figure 4. ZF and ZF-Res](image)

From figure 4, it can be seen that the driver detection using ZF-Res network is better than ZF for the driver's boundary frame, and it is better to capture the local details of the driver's bad behavior.
Table 2. Comparison between the single BP neural network and the cascade classifier

|       | Proposals | Average detection time/(s/frame) |
|-------|-----------|----------------------------------|
| ZF    | 300       | 0.049                            |
| VGG-16| 300       | 0.092                            |
| ZF-Res| 300       | 0.053                            |

Compared with ZF and ZF-Res, the detection effect of ZF-Res network is better when the detection time of single frame is similar.

4. Conclusion
In this paper, a driver detection algorithm based on improved Faster R-CNN is proposed based on RGB images captured by vehicle-mounted cameras. The residual structure is introduced into the ZF network to design the ZF-Net network, in order to improve the accuracy and maintain the real-time performance. LRN (Local Response Normalization) is replaced by BN (Batch Normalization), which simplifies parameter adjustment and accelerates network convergence. Experiments on a self-built driver image database demonstrate the effectiveness of the driver detection method.

5. Acknowledgment
The authors would like to thank the editor and the anonymous reviewers for their valuable comments and constructive suggestions. This work was supported by the National Natural Science Foundation of China (No.61374194), National Key Science & Technology Pillar Program of China (No.2014BAG01B03), and Key Research & Development Program of Jiangsu Province (No. BE2016739), and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

6. References
[1] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2014:580-587.
[2] Girshick R. Fast R-CNN [J]. Computer Science, 2015.
[3] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]// International Conference on Neural Information Processing Systems. MIT Press, 2015:91-99.
[4] Lecun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436-444.
[5] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 779-788.
[6] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks [J]. Computer Science, 2014, 8689: 818 - 833.
[7] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. Computer Science, 2014: 1-14.
[8] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2015: 1-9.
[9] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2016: 770-778.
[10] Huang Gao, Liu Zhuang, Weinberger K Q, et al. Densely connected convolutional networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2017.
[11] Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [J]. 2015:448-456.
[12] Glorot X, Bengio Y. Understanding the difficulty of training deep feed forward neural networks [J]. Journal of Machine Learning Research, 2010, 9: 249-256.