Deep Convolutional Neural Network Based Traffic Vehicle Detection and Recognition

Yukun Rao, Guanwen Zhang(✉), Wei Zhou, Changhao Wang, and Yu Lv

Northwestern Polytechnical University, Xi'an 710072, China
{guanwen.zh,zhouwei}@nwpu.edu.cn

Abstract. Traffic vehicle detection and recognition is a core technology of advanced driver assistant system (ADSD) for the intelligent vehicle. In this paper, we employ the convolution neural network (CNN) to perform the end-to-end vehicle detection and recognition.

Two vehicle classification CNNs are proposed. One is a convolution neural network consisting of four convolution layers and another is a multi-label classification network. The first networks can achieve the accuracy more than 95% while the second can achieve the accuracy more than 98%. Due to the multiple constraints, the proposed multi-label classification network is able to converge fast and achieve higher accuracy.

The vehicle detection model proposed in this paper is a model on the basis of the network model single shot multibox detector (SSD). Our network model employs the network proposed for vehicle classification as a basis network for feature extraction and design a multi-label loss for detection. The proposed network structure can achieve 77.31% mAP on the vehicle detection dataset. Compared with that of SSD network model, the obtained mAP is improved by 2.17%. The processing speed of proposed vehicle detection network can reach 12FPS, which can meet the real-time requirements.

Keywords: Intelligent vehicle · traffic vehicle classification · Multi-label classification · Traffic vehicle detection

1 Introduction

Vehicle detection and recognition is an important research course in the field of vision for unmanned (intelligent) driving technology, it is also a core technology of unmanned driving. In unmanned driving system, vehicle detection and recognition is the premise and foundation for intelligent vehicles to follow other vehicles, change lane, overtake, obstacle avoidance and other acts. The accuracy

Sponsored by the Seed Foundation of Innovation and Creation for Graduate Students in Northwestern Polytechnical University, ZZ2019140 and National College Students' innovation and entrepreneurship training program, 201810699167.

B. Li et al. (Eds.): IoTaaS 2019, LNICST 316, pp. 427–438, 2020.
https://doi.org/10.1007/978-3-030-44751-9_36

and complexity of vehicle detection and recognition directly affect the overall efficiency and performance of unmanned driving. Vehicle detection and recognition during driving can not only meet the real-time demand, but also can give some of instructions to operate the vehicle. The direction of vehicles and the type of vehicles are two important factors that affect the intelligent vehicle autonomous driving when detecting vehicles.

With the development of computer hardware and the emergence of deep learning, a new idea appears for vehicle detection and recognition in unmanned technology. Compared to the traditional methods which need manual image features extraction, the deep learning ways have better adaptability. Deep learning uses neural network for feature extraction, trains the network layer by layer, shares weights and it can enhance the speed of the image processing operations. Deep learning is mainly used in speech recognition, image recognition and so on. While dealing with these problems, the traditional feature extraction methods have limited ability to express features, but the deep learning can breakthrough these restrictions and meet the needs of computing.

Deep learning is to understand the information contained in images, sounds, and text. Deep learning has many successful cases in the face recognition [9], multi-scale image classification [1,2,6,8], object detection [12] and vehicle recognition [4,7]. Therefore, applying deep learning to the vehicle detection and recognition tasks has very significant meanings.

2 Related Works

Research on vehicle recognition mainly includes two aspects: one is the vehicle detection, another is the vehicle classification. Current vehicle detection is mainly divided into two methods, a single frame image-based method and video-based method. In this paper, we mainly study the image-based vehicle detection method. The early image-based vehicle detection methods mostly base on the appearance of the vehicle, using the edge features of image and the symmetry features to carry out vehicle position. These methods also use HOG [3] characteristics of vehicles to locate the vehicle.

Deformable parts model (DPM) [5] is mainly based on the improvement of HOG characteristics. It uses SVM [13] for classification, but needs multi-angle shooting targets while training. In 2014, Ross Grishick applied the convolution neural network to object detection [10]. In the case of insufficient dataset, pre-training of the network was required while fine-tuning the specific parts of the network. In 2015, his proposed DPM model and convolution neural network (CNN) became two widely used visual tools.

In the literature [11], Sarfraz proposed the use of the shape histogram features of the forward vehicle, and then classifies the images with the minimum distance classifier to achieve the purpose of quickly identifying the vehicle type. In the literature with the input image feature extraction, the extracted features will be compared to the characteristics in the database to find the smallest picture, and this picture corresponds to the model of the input image. In 2015, Zhang

Hongbing and others used the vehicle's HOG characteristics to position the vehicle, and then the type of vehicle was classified. His method not only reduced the computational complexity while improving the speed of calculation. 2015, Zhang [14] and others proposed a vehicle type recognition method based on convolution neural network.

In conclusion, the traditional method to vehicle detection and recognition has some disadvantages. It has huge amount of computation and pre-prepared work, besides the accuracy and detection speed can not meet the requirements of real-world application. The method based on deep learning has high accuracy and faster speed comparing to traditional ways. Besides, it does not need a lot of human pre-prepared works. So in our paper, we use deep learning ways to realize vehicle detection and recognition. On the one hand, we use tiny neural network to realize the detection task. Comparing to SSD network, we can save more time while training and increase the detection speed. On the other hand, our model can meet the requirements of multi-label task. Since in vehicle detection task, we should take vehicle type and orientation into considerations, our model can training two label at the same time. From the experiment results, we can see that with limitation of two labels, the network can converge faster and reach a higher accuracy.

We will discuss our model from following part:

Constructing CNN Network Model of Vehicle Classification: Design and train CNN network model to realize the classification task of multiple types of vehicles in CompCars (Comprehensive Car Dataset), and realize multi-label vehicle classification task. Two types of tags are trained on a network model and the training structure is compared with the training results of single labels.

Construction of Vehicle Detection and Recognition Network Model: Based on the above mentioned training model, the network structure is expanded by increasing the convolution layer, and the convolution layer Regression attribute (Regression Loss) of the candidate area rectangle is defined, to achieve an end-to-end network architecture model for vehicle detection and recognition tasks.

Do Optimized Training of Vehicle Detection and Recognition Network Model: Using the network model obtained in training in (1), the new CNN network model is initialized fine-tuned and trained, and by adjusting the network structure and training parameter the network model is further optimized to obtain higher detection and recognition accuracy and precision.

3 Proposed Approach

3.1 Constructing CNN Network Model of Vehicle Classification

For vehicle classification, we use convolution neural network which contains four convolution layer defined as four-layer convolution neural network. The four-layer

convolution neural network is mainly composed of convolution layer, pooling layer, ReLU layer, full connected layer, accuracy layer and loss layer. Its structure is shown in Fig. 1.

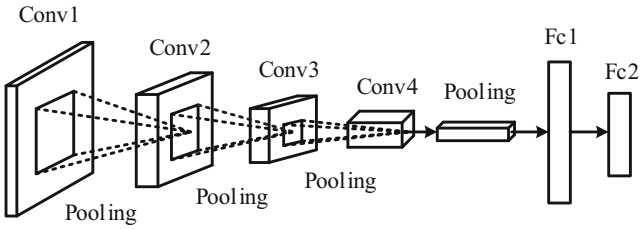

Fig. 1. Four-layer convolution neural network.

The multi-label classification network structure is obtained by modifying the structure of four convolution neural network. The network can simultaneously train a dataset with two labels. On the basis of the original four-layer convolution neural network, the multi-label classification network adds a data layer, an accuracy layer and a loss layer. The new data layer is to give the same images with another set of labels. The new accuracy layer and loss layer is to calculate the network classification accuracy and loss for the second category. Figure 2 shows the structure of the multi-label vehicle classification network model.

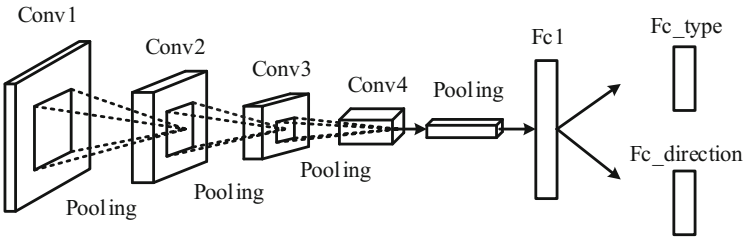

Fig. 2. Multi-label vehicle classification network model.

The accuracy layer of the network is used to calculate the accuracy which is defined as Eq. (1). The accuracy is a rate of the correct predictions and the total number of labels.

$$Ac = \frac{M}{N} \times 100\% \tag{1}$$

In Eq. (1), Ac is the classification accuracy; M indicates the number of correct labels; N indicates the total number of labels.

The loss layer of network is used to calculate classification loss. We use Softmax With Loss function which is combined by two calculation process. Softmax function $sigma(z) = (\sigma_1(z), ..., \sigma_m(z))$ is defined as Eq. (2).

$$\sigma_i(z) = \frac{\exp(z_i)}{\sum_{j=1}^{m} \exp(z_j)}, i = 1, ..., m \tag{2}$$

The next operation is the calculation of the Multinomial Logistic Loss defined in caffe, shown as Eq. (3).

$$l(y, o) = -\log(o_y) \tag{3}$$

We can obtain the Equation of Softmax With Loss by combining Eqs. (2) and (3). It is expressed by Eq. (4).

$$l(y, z) = -\log(\frac{e^{z_y}}{\sum_{j=1}^{m} e^{z_j}}) = \log(\sum_{j=1}^{m} e^{z_j}) - z_y \tag{4}$$

where m refers to the number of labels on current dataset, y means current label. This classification model is designed as a basis of vehicle detection and recognition model.

3.2 Construction of Vehicle Detection and Recognition Network Model

Vehicle detection model is mainly divided into two parts, including vehicle positioning and vehicle classification. Therefore, the most basic network structure is a vehicle classification network structure combined with a vehicle positioning network structure, the structure diagram shown in Fig. 3.

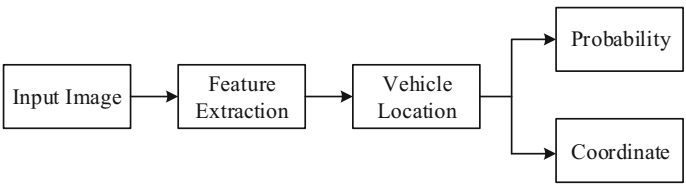

Fig. 3. The basic structure diagram of vehicle detection model.

The vehicle detection network is based on the four-layer convolutional neural network, which can generate a series of fixed-size bounding boxes on the different layers of the feature map. For each bounding box, it is necessary to determine whether it has a target and whether it is a fraction of a target, and finally a non-maximal suppression is added to produce the final test result. In each feature map of the classification network, a series of bounding box is defined. The bounding box is linked with each other in form of convolution. On each unit

of the feature map, we predict each compensation value related to the bounding box shape, and assume that the image in each bounding box is the score of each object. Finally, we find out the highest score of the bounding box, and the position coordinates of the box is the position of the vehicle.

As shown in Fig. 4, for each picture, the network predicts whether the input image contains vehicle and decides the direction and type of vehicle. If it contains the vehicle in the image, the predicted vehicle coordinates should be also given at the same time. The last layer of the network is connected to two loss layers like the multi-label classification model. Each consists of two parts, including the confidence loss and the target prediction position which is the regression loss related to its true position.

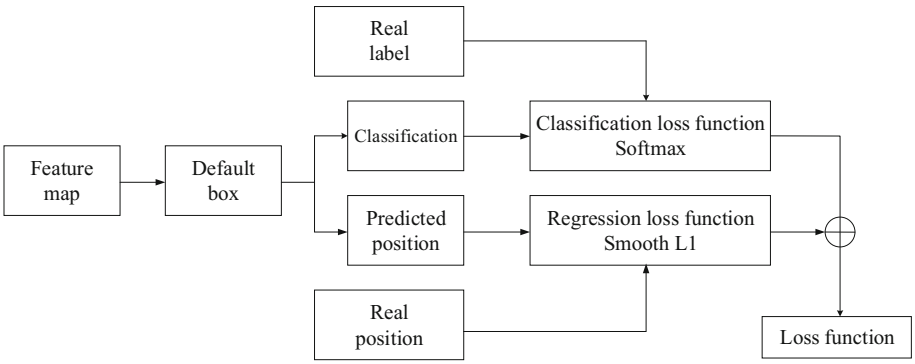

Fig. 4. The principle of vehicle detection model.

Assume that m feature maps are used for prediction, the size of default box in each feature map is calculated as Eq. (5).

$$s_k = s_{k-1} + 15 \times (k-1), k \in [2, m] \tag{5}$$

s_1 represents the smallest size of default box in the smallest feature map between the given feature map.

Then expand the default box into different shape so that it can predict different shape cars. At each location of the feature graph, the network uses multiple bounding boxes with different aspect ratios as $\alpha_r = \{1, 2, 3, 1/2, 1/3\}$, The width and height of the bounding box on different layers are shown in Eq. (6).

$$w_k^\alpha = s_k\sqrt{\alpha_r}, h_k^\alpha = \frac{s_k}{\sqrt{\alpha_r}} \tag{6}$$

When $\alpha_r = 1$, an additional frame is added, and the aspect ratio is calculated as shown in Eq. (7).

$$s_k^{,} = \sqrt{s_k s_{k+1}} \tag{7}$$

Finally, for each position on the feature map, there are six different sizes of bounding boxes for matching. The center of each bounding box (m, n) is shown in Eq. (8).

$$m = \frac{i + 0.5}{|f_k|}, n = \frac{j + 0.5}{|f_k|} \tag{8}$$

Where $|f_k|$ is the size of k_{th} square feature map $[0, |f_k|)$. At the same time, mapping the bounding box coordinates to [0,1] range.

3.3 Define the Loss Function of Multi-label Vehicle Detection and Recognition Network

As we all know, the type and direction of other cars are two important elements to decide which driving order should be executed while driving. So the loss of multi-label vehicle detection and recognition network is combined by two parts. One is the vehicle type detection part and the other is the vehicle direction detection part. We define the loss of vehicle type detection as $error_{type}$ and the loss of vehicle direction detection as $error_{dir}$. The total loss of the whole network is defined as Eq. (9).

$$loss = \alpha error_{type} + \beta error_{dir} \tag{9}$$

where α and β are influence factors of the two kinds of loss.

We set $\alpha = 0.5$ and $\beta = 0.5$ while training our multi-label vehicle detection and recognition network, which means the loss of vehicle type detection and vehicle direction detection have the same effect on the final loss.

By combining two loss, the network performs better in vehicle detection. Since there are two kinds of loss to restrict the network while training the network, the network can converge more quickly and it would be harder to diverge.

4 Experiment Results

4.1 Experiment Result of Vehicle Classification

We use the four-layer convolution neural network model to train and test the vehicle classification datasets. Vehicle classification datasets are divided into two types. One is orientation of vehicle and another is type of vehicle. In the vehicle's orientation dataset, there are 3200 front vehicle images, 3200 rear vehicle images and 3200 background images for training; 800 front vehicle images, 800 rear vehicle images and 800 background images for testing. In the vehicle's type dataset, there are 1600 car images, 1600 minibus images, 1600 bus images, 1600 truck images, 3200 backgrounds for training and 400 car images, 400 minibus images, 400 bus images, 400 truck images, 800 background images for testing.

For vehicle orientation dataset, the loss is 0.081 for final test and the accuracy is 97.75%. The accuracy of each label is shown in Table 1. For vehicle type dataset, the loss is 0.082 for final test and the accuracy is 98.04%. and the accuracy of each label is shown in Table 2.

Table 1. The accuracy of each label in vehicle orientation dataset for four-layer CNN.

Type	Total number	Correct number	Accuracy
Front car	800	792	99.00%
Rear car	800	796	99.50%
Background	800	758	94.75%
Total	2400	2346	97.75%

Table 2. The accuracy of each label in vehicle orientation dataset for four-layer CNN.

Type	Total number	Correct number	Accuracy
Car	400	393	98.35%
Minibus	400	395	98.50%
Truck	400	397	99.25%
Bus	400	398	99.50%
Background	800	771	96.38%
Total	2400	2353	98.04%

Then, the multi-label classification network model is used to train and test the vehicle classification datasets. For vehicle orientation dataset, the loss is 0.012 for final test and the accuracy is 97.91%. The accuracy of each label is shown Table 3.

For vehicle type dataset, the loss is 0.072 for final test and the accuracy is 98.34%. The accuracy of each label is shown in Table 4.

Comparing the data in Tables 1, 2, 3 and 4, we find that the multi-label classification network model can reach higher accuracy than four-layer convolution neural network in vehicle classification dataset. The accuracy increases by around 0.3%. Besides, multi-label classification network model performs better than four-layer convolution neural network. But we still need to improve our model because the accuracy of background is still too low. Due to the limitation of two series of label, multi-label classification model can converge faster and obtain higher accuracy.

Table 3. The accuracy of each label in vehicle orientation dataset for multi-label classification network model.

Type	Total number	Correct number	Accuracy
Front car	800	793	99.13%
Rear car	800	791	98.88%
Background	800	766	95.75%
Total	2400	2350	97.91%

Table 4. The accuracy of each label in vehicle type dataset for multi-label classification network model.

Type	Total number	Correct number	Accuracy
Car	400	398	99.35%
Minibus	400	396	99.00%
Truck	400	396	99.00%
Bus	400	395	98.75%
Background	800	775	96.88%
Total	2400	2360	98.34%

4.2 Experiment Result of Vehicle Detection

The four-layer convolution neural vehicle detection network is based on the four-layer convolution neural network. First, to extract the feature by using the four-layer convolution neural network in previous experiment. Then, selecting different feature layers to set the bounding box. For each bounding box, the calculation of the position error and the category error ultimately determine the position of the vehicle in the picture. When using the basic network, it is necessary to modify the basic network. The last two fully connected layers of the network are all transformed into a convolution layer. Through this operation, all the final connection layer data can be transformed into a feature map for presetting the bounding box. At the same time, to delete the original network loss layer and accuracy layer.

We select the conv4 layer, fc1 layer and the fc2 layer to set the bounding box. The bounding box size is set, as shown in Table 5. The network is trained on the vehicle test set. The training set consists 5639 images while the test set consists 1377 pictures. The final test results of the network mAP is 72.39%.

Table 5. Four-layer convolution neural network's bounding box parameter setting.

Layer	Size of feature map	The original map area	The size of bounding box
Conv4	38×38	8×8	$30 \times 30, 42 \times 21, 21 \times 42$
Fc1	19×19	16×16	$60 \times 60, 83 \times 83, 84 \times 42$ $42 \times 84, 104 \times 35, 35 \times 104$
Fc2	19×19	16×16	$60 \times 60, 83 \times 83, 84 \times 42$ $42 \times 84, 104 \times 35, 35 \times 104$

But for multi-label network, we expand the four-layer convolution neural vehicle detection network with another MultiBoxLoss layer to train two kinds of label of the vehicles. One mulitiBoxLoss layer is used to train vehicle orientation

set while another is used to train vehicle type set. The bounding box setting is the same as the four-layer convolution neural vehicle detection network.

When testing the multi-label network, we first obtain the detection results of vehicle orientation. Then comparing the detection results of vehicle type with the obtained results, find the vehicle type of the bounding box where vehicle orientation is ensured. The test results of the multi-label network mAP is 77.39%. Figure 5 shows some of the picture test results.

Fig. 5. Testing results of multi-label vehicle detection network.

The comparing result is shown in Table 6. It can be concluded that the SSD network can achieve good vehicle detection effect for the data set recorded by the car camera, but the simple convolution neural network structure can achieve the same vehicle detection effect, while spending less on time costs. In the situation of the same detection effect, the four-layer convolution neural network constructed in this paper can identify the image faster because of its simple network structure and lower complexity calculation. The mAP value recognized by ResNet18 network is not high, and the reason is there are more error recognition objects. At the same time, as ResNet18 network structure is deep, the calculation complexity of the network weight updating is greater, resulting in the longer time taken to detect the picture.

Table 6. The accuracy of each label in vehicle type dataset for multi-label classification network model.

Basic network	mAP(%)
Four-layer convolution neural network	72.39
Four – layerconvolutionneuralnetwork (Multi – label)	**77.31**
VGG16(SSD)	75.14
ResNet18	74.24

In conclusion, to detect objects in the case of different situations, the network layer is not the deeper the better, but it depends on the contents of the dataset and detection requirements. If we need faster detection speed, we can use the lower convolution neural network to achieve detection. Moreover, if we need a higher accuracy, then a deeper network can be used to achieve our goals. Besides, training network with multi-label tasks can make the network work better in each task.

5 Conclusion

This paper mainly discuss about the multi-label vehicle detection network. First, a vehicle classification network is needed and it is designed as a basis network of vehicle detection and recognition network. We use four-layer convolution neural network as classification network. Small size network can reach the same accuracy as VGG net but has a higher speed than it. Next, we expand the classification network and obtain a multi-label classification network in order to meet the needs while driving. Last, we expand the multi-label classification network and obtain the multi-label vehicle detection network. From the results, training two categories on one network at same time can not only save time, but also make the network converge faster and achieve higher accuracy. So for multi-task detection, multi-label network can be considered to achieve better result.

References

1. Ciregan, D., Meier, U., Schmidhuber, J.: Multi-column deep neural networks for image classification. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3642–3649, June 2012
2. Ciresan, D.C., Meier, U., Masci, J., Gambardella, L.M., Schmidhuber, J.: Flexible, high performance convolutional neural networks for image classification, pp. 1237–1242, July 2011
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 1, pp. 886–893, June 2005
4. Donahue, J., et al.: DeCAF: a deep convolutional activation feature for generic visual recognition, pp. 647–655 (2014)
5. Forsyth, D.: Object detection with discriminatively trained part-based models. Computer **47**(02), 6–7 (2014)
6. Hu, Y., et al.: Algorithm for vision-based vehicle detection and classification, pp. 568–572, December 2013
7. Krause, J., Stark, M., Deng, J., Li, F.F.: 3D object representations for fine-grained categorization. In: 2013 IEEE International Conference on Computer Vision Workshops, pp. 554–561, December 2013
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Neural Inf. Process. Syst. **25**, 1097–1105 (2012)
9. Lawrence, S., Giles, C.L., Tsoi, A.C., Back, A.D.: Face recognition: a convolutional neural-network approach. IEEE Trans. Neural Networks **8**(1), 98–113 (1997)

10. Ren, S., He, K., Ross, G., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 91–99 (2015)
11. Sarfraz, M.S., Saeed, A., Haris Khan, M., Zahid, R.: Bayesian prior models for vehicle make and model recognition, p. 35, January 2009
12. Shams, F.: Joint deep learning for car detection, December 2014
13. Vojislav, K.: Learning and soft Computing: Support Vector Machines, Neural Networks and Fuzzy Logic Models. MIT Press, Cambridge (2001)
14. Zhang, F., Xu, X., Qiao, Y.: Deep classification of vehicle makers and models: the effectiveness of pre-training and data enhancement. In: 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO), pp. 231–236, December 2015