International Core Journal of Engineering Volume 6 Issue 3, 2020
ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0022

Improved Faster RCNN Approach for Vehicles and Pedestrian Detection

Ping Chen, Dan Yu

Logistics Engineering College, Shanghai Maritime University, Shanghai 201306, China.

payne1202@163.com

Abstract

The detection of vehicles and pedestrians is very important in automatic driving and intelligent assi-sted driving. Traditional detection methods require artificial design features for extraction, which has the problem of poor robustness. Faster-RCNN object detection algorithm is applied to vehicle and pedestrian det- ection. The K-Means clustering algorithm is used to cluster the ground truth to determine the length-width ratio of the anchor frame. Design a feature fusion module to improve the algorithm's ability to use shallow features. Compared with the original Faster-RCNN algorithm, MAP improved by 4.78% in comparison test of KITTI data set, and the detection effect of small targets was improved to some extent.

Keywords

Object detection; Deep learning; Faster-RCNN; K-Means; Feature fusion.

1. Introduction

With the acceleration of urban development and the rapid improvement of people's quality of life, car ownership is increasing and road traffic pressure is also increasing. Therefore, autonomous driving and intelligent assisted driving have become popular research topics. During the driving process of a car, detecting vehicles and pedestrians in the driving environment as accurately and quickly as possible is particularly important for automatic driving and intelligent assisted driving.

Vehicle and pedestrian detection belong to the category of object detection. Object detection not only recognizes the type of target object in the picture, but also locates the position of the target object in the picture, and draws a bounding box around the target object for calibration. Traditional object detection methods are mainly composed of three parts: region selection, feature extraction and classifier classification. For the region selection part, the sliding window strategy is most commonly used. Selecting windows of different sizes and sliding on the picture in a certain order to traverse the image to select candidate frames. For the feature extraction part, the extracted features are generally different for different detected objects. In this part, it is necessary to manually design features for different detection objects. Commonly used features in this part are SIFT, HOG [1], etc. As for the classifier part, the feature vector is usually combined with classifiers such as SVM [2] and Adaboost after feature extraction. Traditional object detection algorithms have their limitations. Sliding window strategies require windows of different sizes to traverse the entire picture. This process is time-consuming, produces a large number of object detection frames, and has low detection accuracy. Manually designing features for feature extraction is time-consuming and labor-intensive, and different features need to be designed for different detection targets, which is less robust.

In recent years, deep learning-based object detection technology has developed rapidly. Compared with traditional object detection methods, deep learning-based object detection methods use convolutional neural networks instead of traditional hand-designed features to extract the features of

ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0022

target objects, which can better adapt to the characteristics of detection object diversity, and is more robust [3-9]. At present, deep learning-based target detection methods are mainly divided into two categories: one is a two-stage method, which divides object detection into two major steps. First, it uses the region proposal algorithm to generate candidate regions that may contain targets, and then uses convolutional neural networks to classify and regress the candidate regions to obtain the final detection frame. Such methods mainly include SPP-Net [10], Fast-RCNN [11], Faster-RCNN [12], etc. The other is the one-stage method. The biggest difference from the previous method is that it maps the window set in the original image directly to the feature map generated by the convolutional neural network, and returns the category of the window by deep features. And position offset to finally get the object detection frame. These methods mainly include YOLO [13], SSD [14], YOLOv3 [15], and so on.

In this paper, Faster-RCNN object detection algorithm is used to detect vehicles and pedestrians. First, the K-Means clustering algorithm is used to cluster the target frames to determine the aspect ratio of the target frames. Modified the size of the anchor frame for the relatively small size of vehicles and pedestrians in the distance. Then improve the feature extraction network. The features of the shallow and deep features are fused and then object classification is performed to improve the algorithm's use of shallow features in object detection, and the improved algorithm is verified on the KITTI dataset.

2. Vehicle and pedestrian detection method network structure

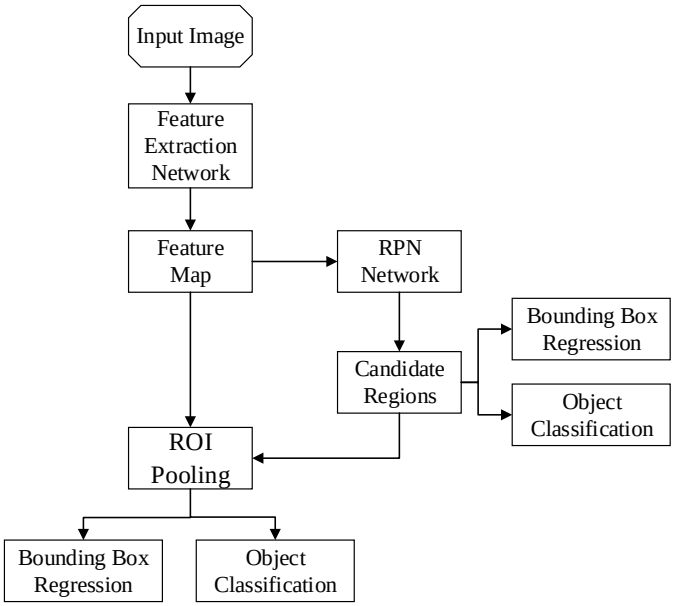

Fig.1 Vehicle and pedestrian detection method network structure

This paper uses Faster-RCNN object detection algorithm to detect vehicles and pedestrians. The network structure, see Fig. 1. After preprocessing, the input image is sent to a feature extraction network (a deep convolutional neural network) for feature extraction. The picture obtained by the last convolution operation is called a shared feature map. Mainly used for RPN network and ROI pooling layer. The shared feature map is then sent to the RPN network. The RPN network first performs a 3x3 convolution operation on the shared feature map, and then uses each pixel point on the obtained feature map as an anchor point, and generates a series of regional suggestion boxes according to the configured size and length-width ratio. The region suggestion frame performs the classification of the foreground and background, and the bounding box regression. The position of the bounding box of the ROI is corrected for the first time to obtain a certain number of candidate regions. The filtered candidate regions is mapped onto the shared feature map to select the feature map, and the selected feature map is pooled by the ROI to output a fixed-size feature map. Finally, the fully-connected layer is used to classify the target frame into the target category. And the accurate

ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0022

regression of the target frame is performed to determine the position information of the detection object.

3. Improve Faster-RCNN network structure

3.1 Improve RPN network

The input of the RPN network is an image of any size, and the output is a series of candidate regions. Each candidate region corresponds to the probability of the existence of a target and the deviation information of the target position.

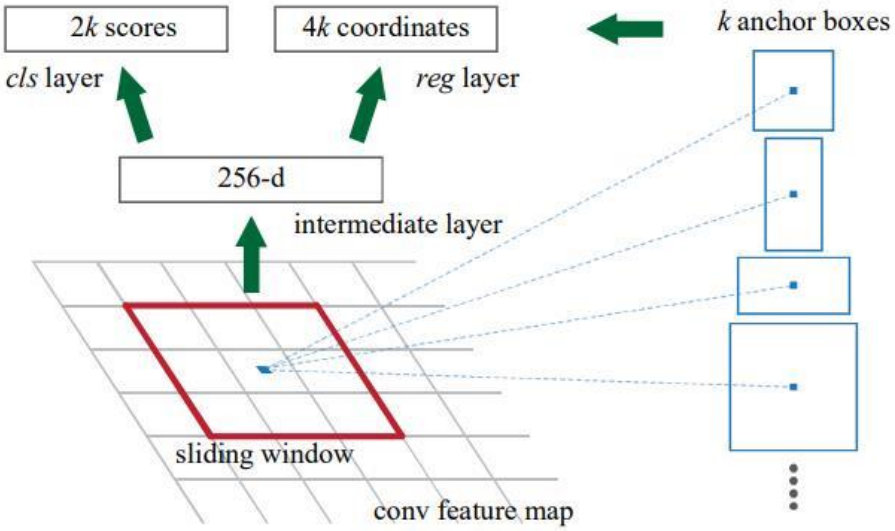

Fig.2 RPN Network

The structure of the RPN network, see Fig. 2. The feature map of the convolution layer here is the shared feature map obtained after the input image passes the feature extraction network. After the shared feature map is sent to the RPN network, the RPN network first performs a convolution operation on the shared feature map through a 3x3 size convolution kernel to obtain a feature map with a channel number of 256. The size is the same as the shared feature map. Each pixel point on the feature map after the convolution operation is used as an anchor point, and is mapped to the coordinate point position corresponding to the original picture. Each point corresponds to 9 different rectangular windows. These 9 different rectangular windows are composed of three aspect ratios {1: 2, 1: 1, 2: 1} and three sizes $\{128^2, 256^2, 512^2\}$.

This paper uses K-means clustering algorithm to cluster the target frames in the dataset to determine the aspect ratio of the anchor frame. The Euclidean distance will make the target frame with a larger size produce an error larger than the target frame with a smaller size. We hope that the size of the error has nothing to do with the size of the target frame. Therefore, the distance formula here draws on the distance formula of the anchor box in YOLO algorithm, such as Formula (1) shows:

$$d(box, centroid) = 1 - IOU(box, centroid) \quad (1)$$

Finally, the K-means clustering algorithm was used to determine the aspect ratio of the anchor box as {2: 5, 3: 2, 2: 1}. At the same time, the anchors were modified for the small size of pedestrian objects and distant vehicle objects in the dataset. The size of the box is $\{32^2,64^2,128^2,256^2\}$. After the improvement, each anchor point generates a set of 12 different candidate frames. The number of anchor frames has been increased from the original 9 to 12. More anchor frames improve the detection accuracy. At the same time, the size of the anchor frame is much smaller than before. Compared with the modification, the reduction of the anchor frame size is reduced. The calculation amount is reduced, and the calculation time is shortened.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0022

3.2 Feature Fusion

The feature extraction network in Faster-RCNN object detection algorithm is VGG16 convolutional neural network. VGG16 convolutional neural network has 13 convolutional layers. A non-linear activation function Relu (The Rectified Linear Unit) is added after each convolution layer. VGG16 is divided into 5 convolution modules, namely Conv1_2, Conv2_2, Conv3_3, Conv4_3, Conv5_3. Each of the first 4 groups is followed by a Max Pooling layer for dimension reduction. After the Conv5_3 convolution operation, the shared feature map will be obtained, and the shared feature map will be sent to the RPN network and ROI layer for calculation.

The original Faster-RCNN network obtained the feature map of the input picture after 13 convolutional layers. This process well extracted the deep information of the picture, but it also lost the shallow information of the picture. As the convolutional layer deepens, the receptive field also continues to increase, which will be difficult to detect for smaller objects in the picture, such as distant vehicles and pedestrians. Aiming at this problem, this paper proposes a feature fusion method to increase the shallow level information in the shared feature map, see Fig. 3, the feature map of the Conv1_2 layer is reduced to the same size and the same number of channels as the feature map of the Conv5_3 layer. This process is implemented by three convolution operations. The convolution kernel size of the three convolution operations is 3x3, and the step size is set to 2. The number of channels after the first convolution is 128, the number of channels after the second convolution is 256, and the number of channels after the third convolution is 512. Finally, the feature map obtained is fused with the feature map of the Conv5_3 layer.

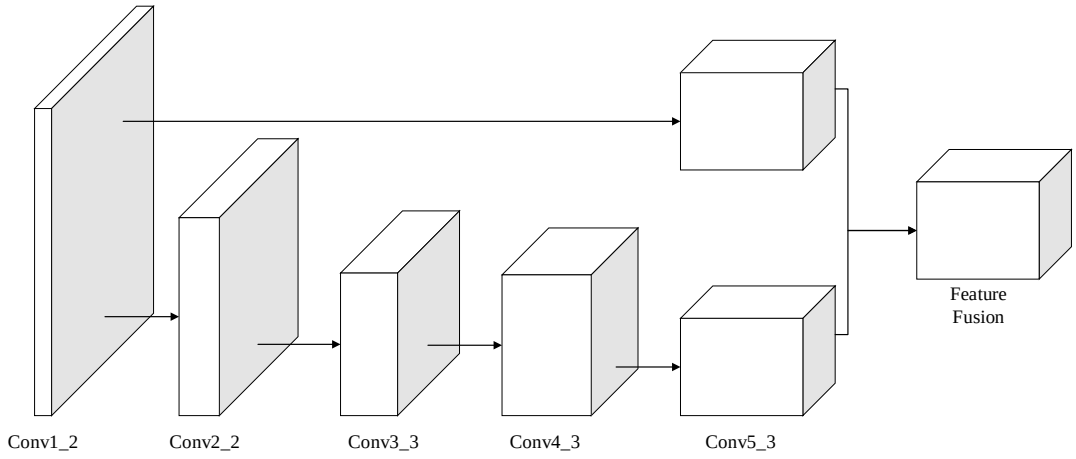

Fig. 3 Feature Fusion

The feature fusion here uses the feature map summation method, and the calculation method of the feature map summation is shown in formula (2):

$$Z_{add} = \sum_{i=1}^{C} (X_i + Y_i) * K_i = \sum_{i=1}^{C} X_i * K_i + \sum_{i=1}^{C} Y_i * K_i \quad (2)$$

4. Experimental analysis and results

4.1 Data set and experimental platform

This article uses the public KITTI data set. Download the KITTI dataset and convert it to the PASCAL VOC dataset type required for Faster-RCNN algorithm training. Then ignore the “DontCare”, “Misc”, and “Cyclist” classes that are too small in size and have inaccurate labeling information. Merge the 'pedestrian' and 'pedestrian (sitting)' classes into one class, collectively called the 'pedestrian' class, merge the 'car', 'van', 'truck', 'tram' classes into one class, collectively called 'car' class. A total of 7481 labeled images in the KITTI data set are divided into a training set, a validation set, and a test set according to a ratio of 8: 1: 1.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0022

The training data set image is flipped left and right to expand the data set, and then sent to the detection network model in this article for training. The fine-tuning strategy currently used in deep learning is used to select the pre-trained model on the Image Net classification task to initialize the training network.

On the Ubuntu 18.04-LTS operating system, the Tensor flow deep learning framework was used to train on a computer equipped with 16GB of memory, 8GB NVIDIA GeForce GTX1070 GPU and CUDA10.1 and CUDNN7.5 computing platform.

4.2 Experimental results and analysis

In this paper, the training set data is sent to the model for training and tested on the test set. Recorded vehicle detection accuracy, pedestrian detection accuracy, average detection accuracy, and detection speed before and after model improvement, see Table 1.

Table 1 Test results before and after model improvement

Numble	Model	Car AP(%)	Pedestrian AP(%)	MAP(%)	Time(s)
1	Faster-RCNN	81.40	67.69	74.55	0.1
2	Improved Faster-RCNN	88.45	70.20	79.33	0.091

Compared with the original Faster-RCNN algorithm, the improved Faster-RCNN object detection algorithm has improved the average detection accuracy of the vehicle by 7.05%, the average detection progress of pedestrians has increased by 2.51%, the average detection accuracy has increased by 4.78%, and the detection speed has also been certain. This is because the target frame is clustered by the K-Means clustering algorithm, so that the determined candidate frame aspect ratio is more in line with the actual situation, which can better fit the detection object. In addition, a feature fusion method is added to fuse the feature information of the shallow layer with the feature information of the deep layer, so that the model can use the feature information of the shallow layer in the selection of candidate frames and the classification of the object, thereby improving the detection accuracy.

4.3 Experimental results show

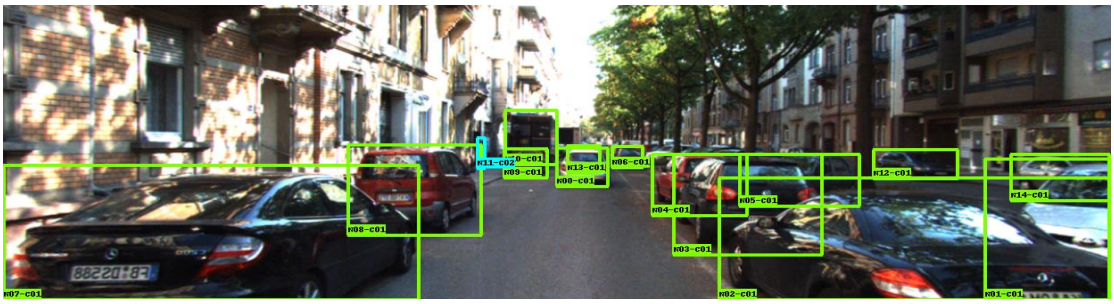

Fig. 4 Improved Faster-RCNN vehicle detection

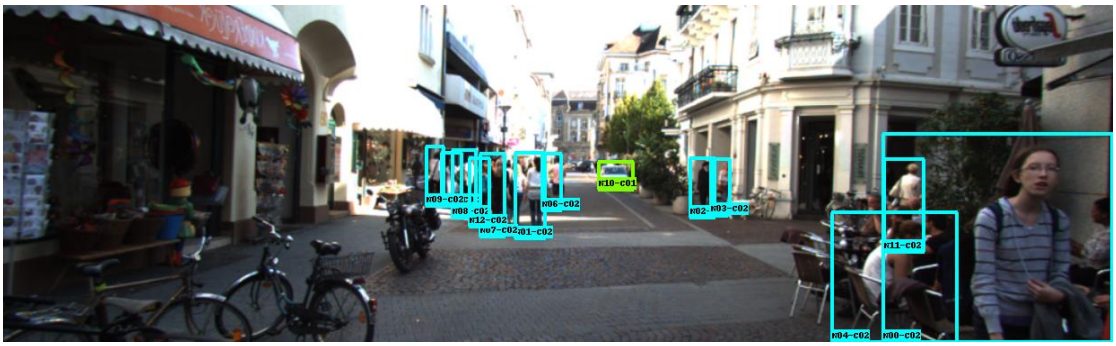

Fig. 5 Improved Faster-RCNN pedestrian detection

The improved Faster-RCNN algorithm has better detection effect on vehicles and pedestrians with smaller size in the picture. This improves the detection accuracy of vehicles and pedestrians, see Fig. 4 and Fig. 5.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0022

5. Conclusion

In order to better detect vehicles and pedestrians in the vehicle driving environment, this paper improves the Faster-RCNN object detection algorithm, and trains and tests on the KITTI dataset. Use the K-Means clustering algorithm to cluster the target frame to determine the length and width ratio of the anchor frame. For smaller pedestrians and smaller vehicles in the distance, improve the size of the anchor frame. The detection accuracy of the model is added to the feature fusion method, which enables the model to use the feature information of the shallow layer of the picture in the selection of candidate frames and the classification of objects, thereby improving the overall target detection accuracy of the model.

Although the method in this paper improves the detection accuracy and speed of vehicles and pedestrians, it still has certain shortcomings. First, the clustering of the target frame by the K-Means clustering algorithm is manually completed before model training. To modify the aspect ratio and size of the anchor box. After the training data set changes, the target box still needs to be artificially clustered to determine the aspect ratio of the anchor box. In addition, although the method in this paper has improved the detection accuracy and speed of vehicles and pedestrians, there is still a lot of room for improvement. These shortcomings need further research and improvement.

References

[1] Dalal N. Triggs B. Histograms of Oriented Gradients for Human Detection[C]. Computer Vision and Pattern Recognition, IEEE Computer Society Conference on. IEEE, 2005, 1:886-893.

[2] Burges C J C.A Tutorial on Support Vector Machines for Pattern Recognition [J]. Data Mining and Knowledge Discovery, 1998, 2 (2):121-167.

[3] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C] //International Conference on Neural Information Processing Systems, 2012: 1097-1105.

[4] Lecun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition [J].Neural Computation, 2014,1(4):541-551.

[5] Li Y, He K. Sun J.R-FCN: object detection via region based fully convolutional networks[C]//Advances in Neural Information Processing Systems. 2016: 379-387.

[6] Shiyu Cao, Yuehu Liu, Xinzhao Li. Vehicle object detection based on Fast R-CNN [J] .Journal of Image and Graphics, 2017, 22 (5): 671-677.

[7] Xiliang Yan, Liming Wang. Handwritten Chinese Character Recognition System Based on Convolutional Deep Neural Network [J]. Computer Engineering and Applications, 2017, 53 (10): 246-250.

[8] 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.

[9] Huansheng Song, Xiangqing Zhang, Baofeng Zheng, Teng Yan. Vehicle object detection in complex scenes based on deep learning methods [J / OL]. [2017- 03-31].

[10] He K , Zhang X , Ren S , et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(9):1904-16.

[11] Girshick R. Fast R-CNN[J]. Computer Science, 2015.

[12] Ren S, He K, Girshick R, et al. Faster R-CNN：towards real- time object detection with region proposal networks[C] //Advances in Neural Information Processing Systems. 2015: 91-99.

[13] 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.

[14] Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector[C] //European Conference on Computer Vision. Cham: Springer, 2016: 21-37.

[15] REDMON J,FARHADI A.YOLO V3:an incremental improvement［C］//IEEE Conference on Computer Vision and Pattern Recognition, 2018:1-6.