Research on Vehicle Object Detection Algorithm Based on Improved YOLOv3 Algorithm

Jin Liu and Dongquan Zhang
School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing, China
Email: 18121410@bjtu.edu.cn

Abstract. Vehicle object detection is one of the important research directions in the field of computer vision. Aiming at solving the problems of low accuracy, slow speed, and unsatisfactory results of using traditional methods to detect the object of the vehicle in front of the driverless car on the road, this paper proposes an improved YOLOv3 vehicle target detection algorithm which we name it F-YOLOv3. First the multi-scale prediction network model is improved according to actual traffic conditions and efficiency requirements based on the original general object detection YOLOv3 algorithm. Then a scale prediction layer is added to improve the detection accuracy of large vehicles and improved k-means++ the algorithm is used to improve the effect of anchor box dimensional clustering and the detection speed. At last an experiment was conducted on a self-made dataset and compared with YOLOv3 in order to test the effectiveness of the F-YOLOv3 algorithm. The test results show that the improved F-YOLOv3 model has a precision mAP of 91.12% and a speed of 59FPS, which are better than the traditional general object detection YOLOv3 algorithm. Therefore, the algorithm has better performance and popularization prospect in vehicle object detection.

1. Introduction
Vehicle object detection is an important branch of computer vision, which is also the foundation of driverless, intelligent transportation, vehicle tracking and other fields. In the current practical application scenarios, the main challenge of vehicle object detection is to explore the relationships of the accuracy, efficiency and actual demand of detection algorithm. How to detect vehicle objects in complex traffic scenes quickly and accurately has been the research fields of computer vision and transportation related interdisciplinary. Therefore, a large amount of effective research has been done by Chinese and foreign scholars and many vehicle object detection algorithms have been successfully developed.

Early traditional vehicle detection algorithms were mainly based on artificial feature extraction. Van [1] et al. used Histogram of Oriented Gradients (HOG) and Support Vector Machine (SVM) method to classify vehicle objects, but their method is complex and the robustness is insufficient as some morphological operations, image processing and other pre-processing must been performed for vehicle images. Hamid [2] et al. improved the Haar features and trained the Adaboost cascade classifier to classify and identify the vehicle, and the reliability of vehicle object detection is improved, but the detection effect is reduced when the relative speed of the vehicle is too fast. The above artificial feature extraction algorithm first extracts candidate regions through sliding window filtering and other methods, then manually extracts features, and finally uses a classifier for classification and recognition. Artificial feature extraction algorithm suffers either from high computational costs, low accuracy, or insufficient robustness.
Convolutional neural network algorithm based on deep learning has appeared in order to improve the disadvantages of the above algorithm. Ren [3],[4] and others have proposed three methods of deep learning: RCNN, fast RCNN and fast RCNN, and each of which is a two stage method. In the first stage, a certain image segmentation algorithm is used to select region of interest (ROI). In the second stage, images are input into convolution neural network for classification and regression operation based on the idea of sliding window. The advantage of the methods is that the image features can be fully extracted to achieve accurate classification and positioning, but it has the disadvantages of slow running speed and low efficiency because of the two stage processes. Joseph [5],[6],[7] and others proposed the YOLO series algorithm, which is an object detection method with simple network structure and higher real-time performance. It directly transforms the classification and positioning of the target into a regression problem. The advantage of this method is that the algorithm runs fast, but the detection accuracy is relatively reduced.

In this paper, an improved YOLOv3 algorithm which we called F-YOLOv3 is proposed to adapt to the detection of large vehicle targets in traffic scenario image. First, the improved k-means++ algorithm is used to cluster, which improves the effect of anchor box dimension clustering, and then a scale prediction layer is add, which improves the multi-scale prediction network model. The experiment show that the algorithm large target can been more effectively detected, recognition accuracy can been improved and network detection time can been reduced.

2. Research on Vehicle Object Detection
YOLO (you look only once) is a popular and general one stage object detection algorithm, which has been developed to the third generation and is called YOLOv3 [7]. The algorithm structure of YOLOv3 is shown in Figure 1. First a standardized image is used as input to the algorithm. Next the image is divided into S×S grids. Then use these grids to generate class probability map, bounding boxes and confidence score. Finally, the object candidate box with confidence and location is actually output on the image. Object recognition, classification and positioning are transformed into regression problems, which is the core idea of the algorithm. Only one convolutional network is used to predict the classes and location of the object, so as to achieve rapid object detection.

Figure 1. YOLOv3 Architecture

Compared with the previous two generations of YOLO algorithm, the YOLOv3 has made some adaptive improvements, including: multi-label classification, different bounding box prediction, multi-scale recognition, etc.
Multi-label classification. A logistic classifier is used to calculate the similarity of specific label, and the binary cross entropy of each label is used to replace the previous as the classification loss instead of the previous mean square error;

Bounding box prediction. Associating A anchor box with score of 1, to obtain the object with the highest degree of overlap of the anchor box than other anchor boxes.

Multi-scale recognition. Using the feature pyramid network (FPN), YOLOv3 predicts three different scale feature maps, and then extracts features from these scale maps to regression the location and classes.

The new CNN feature extractor uses the DarkNet-53 [7] network based on the improved ResNet residual network as a feature extractor, which makes floating-point operations less and faster than before.

Therefore, YOLOv3 has become one of the best representative algorithms for object detection.

2.1. Network Structure

Image with 3 channels and the scale adjusted to 416×416 is used as the input of the YOLOv3 network. First of all, convolution + (Batch Normalization) BN operation is performed on the image of uniform size. Through 32-layer convolution operation with a filter size of 3×3, the output is a 416×416 feature map of 32 channels. Then, the feature map through the DarkNet-53 feature extraction structure, from the 0th layer to the 74th layer, there are 53 convolutional layers, which are composed of a series of consecutive 3×3 and 1×1 convolutional layers. The way of accumulating filters realizes the local feature fusion between different scale feature maps. In order to solve the problem of gradient dispersion or gradient explosion of the network, a ResNet residual network is proposed in YOLOv3 and there are 5 residual blocks. Each residual block is made up of several residual units. By using the residual units, the depth of the network is deeper, the gradient fading is avoided, and the structure size of the input-output model is kept constant. Finally, from layer 75th to layer 105th are the feature fusion layers of YOLOv3 network. Figure 2 shows the network structure of YOLOv3.

Feature Pyramid Networks (FPN) is often used to detect objects of different scales. DarkNet-53 feature extraction network uses the up-sampling and fusion feature pyramid structure of the FPN network [8]. The original image is divided into three scales according to the size of the feature map, and each scale is subdivided into S×S equal grids. The 13×13 down sampling feature map is used to detect larger targets, the 26×26 down sampling feature map is used to detect medium-sized targets, and the 52×52 down sampling feature map is used to detect small targets. The final output is a 75-
channel feature map. Then use the fusion of shallow and deep features to obtain more discriminating deep features. Finally, three bounding boxes are predicted with the assistance of three anchor boxes in each cell. On this basis, regression, classification, and positioning are performed to predict the bounding box coordinates and corresponding confidence scores.

2.2. Improvement of Anchor Box Dimension Clustering Algorithm

In the YOLOv3 algorithm [7], a group of anchor boxes with fixed size are introduced for prediction based on the idea of Faster R-CNN algorithm [3]. There are k initial anchor boxes are obtained by dimension clustering of the height and width of the manually labelled anchor boxes in the dataset through k-means algorithm. For different datasets, YOLOv3 sets 9 initial anchor boxes to obtain 9 clustering results, which are: (10×13); (16×30); (33×23); (30×61); (62×45); (59×119); (116×90); (156×198); (373×326). However, the correlation between general dataset classes and vehicle dataset classes is low. And the shooting angle of the general dataset is quite different from that of the rear of the vehicle. The general dataset is not suitable for real-time vehicle detection.

The k-means algorithm randomly selects k initial clustering centres from the sample set, so that causes a low clustering accuracy. After our research, the improved k-means++ algorithm is selected [9]. The threshold \( \epsilon \) is set to cluster the anchor box, and the next initial cluster centre is more likely to be selected from a relatively far point. Intersection over union (IOU) is the ratio between the prediction result of the anchor box and the intersection area and union area of the ground truth, which has nothing to do with the size of the anchor box. This paper uses IOU as the object cluster analysis of measurement. YOLOv3 uses logistic regression to predict the probability of objects contained in the anchor box. If the overlap rate of the anchor box and the real target frame is greater than any other anchor box, the probability of this anchor box is 1; if the overlap rate of the anchor box and the real target frame is greater than 0.5, but it is not the largest, then ignore this prediction. Therefore, the k-means++ clustering algorithm uses IOU as the spatial distance calculation instead of the Euclidean distance, thereby reducing the error generated by the initial anchor box of different sizes. The improved distance measure is:

\[
D_i(x_j) = 1 - IOU(x_j, c_i)
\]

\( x_j \in X = \{x_1, x_2, ..., x_n\} \) is the size sample of the ground truth;
\( c_j \in \{c_1, c_2, ..., c_n\} \) is the cluster centre size;
\( K \) is the number of anchor boxes.

The clustering objective function is the minimum value of the sum of the distance from each sample to its clustering centre, and the calculation formula is as follows:

\[
J(k) = \min \sum_{i=0}^{K} \sum_{j=0}^{x} D_i(x_j)
\]

The improved k-means++ [9] is used to perform dimensional clustering on the anchor box, as shown in Figure 3 for the improved anchor box dimensional clustering. The contour coefficient method is used to analyse the clustering objective function to select the optimal clustering number k. When k is less than the true value 3, \( J(k) \) will drop sharply; and when k reaches the true value 3, \( J(k) \) stops rapidly and drops sharply, the clustering effect decreases, and as k increases continuously, becoming steady. Therefore, the optimal initial anchor box cluster number of the vehicle dataset is 3, and the height and width are (386×147); (208×193) and (117×265).

Figure 3. Anchor box dimension clustering
Table 1 shows the average ratio of IOU of the anchor box. The F-YOLOv3 of the improved k-means++ is 10% higher than the YOLOv3, and the number of anchor box is reduced on the basis of ensuring a higher average IOU ratio. K-means++ can significantly improve the error of classification results, reduce the calculation time and reduce the consumption of computing resources.

| Anchor box generation method | Number of anchor boxes | Average ratio of IOU(%) |
|-----------------------------|------------------------|------------------------|
| YOLOv3                      | 9                      | 73.56%                 |
| F-YOLOv3                    | 3                      | 82.17%                 |

Table 1. A comparison of anchor box average ratio of IOU.

2.3. Improved Multi-Scale Prediction

The idea of Feature Pyramid Network (FPN) is referenced in YOLOv3, which can extract features from different dimensions in an image. The resolution information of the low-level features and the semantic information of the high-level features are used to identify targets on three feature layers of different scales by fusing the features of different levels with the upper sampling. Figure 4 shows the receptive field of the front image under the three scales of 13×13, 26×26 and 52×52 in the vehicle form process. When the front vehicle object is large, especially for buses and trucks, the device is directly mounted in front of the vehicle, and the receptive field of the output feature map at the scale level of 52×52 corresponds to a very small part of the target. At this time, it will be difficult to predict the vehicle object border using an anchor box to ensure the coverage of the priori frame to the large object. When the vehicle object in front is small, the receptive field of the output feature map at the scale level of 13×13 corresponds to a larger part of the target, and it will be difficult to ensure the coverage of the anchor box to the small object by using the anchor box to predict the vehicle object border.

Figure 4. Vehicle targets feature maps

Most of the vehicle targets in the vehicle detection image are large targets. In this paper, the scale recognition module in YOLOv3 is improved. The original 3-scale recognition is modified by 4-scale recognition, which can ensure that a more accurate anchor box can be allocated to large targets in the smaller feature map, and a more accurate anchor box can be allocated to small targets in the larger feature map. In this paper, three scales of YOLOv3 are modified to four scales. Each cell on each scale uses three anchor boxes to predict three bounding box.

Figure 5. Added the fourth vehicle target feature map
Based on the improvement ideas proposed above, Figure 6 shows the improved F-YOLOv3 network structure proposed in this paper. The input of the network is the 416×416 image to be recognized. After feature extraction, multi-scale recognition is used, and finally the recognition results of 4 scales are output, which are 7×7×18, 13×13×18, 26×26×18 and 52×52×18, respectively. Where the 18-dimensional channel represents the information of using 3 anchor boxes to predict 3 bounding boxes, and each predicted bounding box information includes frame coordinates (x, y), size (w, h), and confidence Degree (c) and the probability of belonging to 4 classes.

![Figure 6. Improved F-YOLOv3 network structure](image)

3. Experiments and Analysis of Experimental Results

3.1. Dataset and Experimental Environment Configuration

Vehicle object detection based on deep learning needs to learn features from data samples, and the dataset must be representative. Existing datasets, such as KITTI, are too large, and there are great differences between foreign and domestic driving environments. In this paper, we use our own dataset, called VEL. The dataset consists of 5000 images. It covering four common types of vehicles, including: car, truck, bus, other, all of which are collected from the Internet. The VEL dataset covers almost all possible imaging variations, such as body parts at different scales, lighting, front and back, background, and obstruction.

The number of original vehicle detection datasets is small, three methods are used to expand the data in this paper. Random image translation: The image is shifted horizontally or vertically -0.25 to 0.25 times; Random Image Scaling: The image is randomly reduced or enlarged 0.5 times; Random Image Mirror Flip: The image is randomly flipped to get a mirror image. Using the above data amplification methods, the VEL dataset is obtained for future experiments in this paper. All images are annotated with vehicle class labels and high-precision bounding boxes.
In order to evaluate the improved F-YOLOv3 algorithm, the experimental conditions were implemented on a computer with a main frequency of 3.4 GHz using the TensorFlow 1.14 environment framework configured with acceleration libraries CUDA9.0 and CUDNN7.6.

3.2. Analysis of Experimental Results

In the experiment, the self-made VEL dataset is used, of which 80% was used as the training set of the network and 20% is used as the test set. All experiments were run on a computer with a primary frequency of 3.4 GHz. The results of our F-YOLOv3 algorithm are compared with those of two other typical universal detection methods, Faster R-CNN [3] and YOLOv3 [7]. For Faster R-CNN algorithm, the VGG training detection model is selected, and the self-made VEL test dataset is used. For YOLOv3, use the Darknet-53 training model and use the same VEL test dataset. The F-YOLOv3 algorithm uses the above improved network model, using the VEL training set and test set. During training, the initial learning rate is set to 0.001, and the learning rate is changed by 0.1 times before 50,000 iterations. The parameters are initialized separately. At the same time, some commonly used data enhancement methods are used to increase the amount of data and improve the robustness of the model, such as hue, saturation, and exposure offset. The results of the vehicle object detection experiment are shown in Figure 8.

The results of each model are recorded according to the relevant evaluation indicators, as shown in Table 2. In terms of detection speed, in vehicle object detection, a detector FPS of 24 is considered a real-time detector. As can be seen from the table, Faster R-CNN performs much worse than the other two. The detection speed of YOLOv3 is 26 FPS, which can achieve the effect of real-time detection. F-YOLOv3 can reach 59 FPS. In terms of accuracy, the mAP of F-YOLOv3 is 14.52% higher than Faster R-CNN and 12.44% higher than YOLOv3. The performance of F-YOLOv3 is significantly better than that of Faster R-CNN and YOLOv3. The detection accuracy of Faster R-CNN is better, but the FPS is 5, which is lower than the requirement. Among the four types of vehicles, small cars can achieve better results, because these vehicles have small shapes and high general recognition, and they have completely different characteristics from large vehicles. In contrast, the recognition performance of large passenger cars is worse than that of other types. The main reason is that the characteristics of large vehicles are smaller than those of large trucks.
Table 2. Detection results of different methods for VEL datasets

| Algorithm    | FPS | Car    | Truck  | Bus    | Other  | mAP(%) |
|--------------|-----|--------|--------|--------|--------|--------|
| Faster R-CNN | 5   | 0.9165 | 0.9025 | 0.8865 | 0.8947 | 90.01  |
| YOLOv3       | 26  | 0.7933 | 0.7920 | 0.7799 | 0.7818 | 78.68  |
| F-YOLOv3     | 59  | 0.9215 | 0.9224 | 0.9018 | 0.8968 | 91.12  |

The F-YOLOv3 algorithm proposed in this paper is based on an anchor box dimension clustering and multi-scale prediction, which is compared with the YOLOv3 algorithm in real-time video test. The result is shown in Figure 9.

Figure 9. YOLOv3, F-YOLOv3 model real-time video detection effect

The improved F-YOLOv3 not only guarantees real-time performance, but also has 12.44% higher accuracy than YOLOv3. The F-YOLOv3 algorithm obtains 91.12% accuracy and 59FPS detection rate, which are 15% and 120% higher than YOLOv3, respectively. The performance of F-YOLOv3 algorithm is further verified. Compared with Faster R-CNN and YOLOv3 improved F-YOLOv3 algorithm, the improved F-YOLOv3 algorithm performs better in vehicle object detection, has higher average recognition accuracy, greatly improves the detection speed, and overcomes the shortcomings of real-time detection method.

4. Conclusion
A novel F-YOLOv3 algorithm based on the YOLOv3 algorithm is proposed in this paper in order to improve the real-time, efficiency and accuracy of large-scale vehicles detections. Firstly, the anchor box dimension clustering algorithm is improved, then the multi-scale prediction network model is improved. The model on the VEL dataset built by this paper is trained and compared with other methods YOLOv3 and Faster R-CNN. Experimental results prove that the improved F-YOLOv3 model has a precision mAP of 91.12% and a speed of 59FPS, which are better than the traditional general object detection YOLOv3 algorithm. Therefore, the F-YOLOv3 algorithm has better performance and popularization prospect in vehicle object detection.

5. References
[1] Van Pham H, Lee B-R. Front-view car detection and counting with occlusion in dense traffic flow. Int J Control Autom Syst. 2015 Oct;13(5):1150–60.
[2] Hassannejad H, Medici P, Cardarelli E, Cerri P. Detection of moving objects in roundabouts based on a monocular system. Expert Systems with Applications. 2015 Jun;42(9):4167–76.
[3] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Anal Mach Intell. 2017 Jun 1;39(6):1137–49.
[4] Girshick R, Donahue J, Darrell T, Malik J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition [Internet]. Columbus, OH, USA: IEEE; 2014 [cited 2020 Mar 11]. p. 580–7.
[5] Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. arXiv:150602640 [cs] [Internet]. 2016 May 9 [cited 2020 Mar 11].

[6] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [Internet]. Honolulu, HI: IEEE; 2017 [cited 2020 Mar 11]. p. 6517–25.

[7] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement. arXiv:180402767 [cs] [Internet]. 2018 Apr 8 [cited 2020 Mar 11];

[8] Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature Pyramid Networks for Object Detection. arXiv:161203144 [cs] [Internet]. 2017 Apr 19 [cited 2020 May 6].

[9] Arthur D, Vassilvitskii S. k-means++: The Advantages of Careful Seeding. 9.