ITC 1/50

Information Technology
and Control
Vol. 50 / No. 1 / 2021
pp. 13-27
DOI 10.5755/j01.itc.50.1.25094

Multi-object Recognition Method Based on Improved YOLOv2 Model

Received 2020/01/16 | Accepted after revision 2021/02/17

http://dx.doi.org/10.5755/j01.itc.50.1.25094

HOW TO CITE: Li, X., Shi, B., Nie, T., Zhang, K., Wang, W. (2021). Multi-object Recognition Method Based on Improved YOLOv2 Model. *Information Technology and Control*, 50(1), 13-27. https://doi.org/10.5755/j01.itc.50.1.25094

Multi-object Recognition Method Based on Improved YOLOv2 Model

Xun Li

School of Electronics and Information, Xi'an Polytechnic University, 58 Shangu Roda, Lintong District, Xi'an City, Shaanxi Province; Phone: +13898231612; e-mail: lixun@xpu.edu.cn

Bernoulli Institute, University of Groningen, Groningen, Nijenborgh 9, 9747 AG Groningen, the Kingdom of the Netherlands; Phone: +31633408399; e-mail: xun.li@rug.nl

Binbin Shi, Tingting Nie, Kaibin Zhang, Wenjie Wang

School of Electronics and Information, Xi'an Polytechnic University, 58 Shangu Roda, Lintong District, xi'an City, Shaanxi Province; phones: +15191859877, +13298318532, +15191418916, +18946004933; e-mails: 734931099@qq.com, 2465363101@qq.com, xihua_0169@163.com, wangwenjie@xpu.edu.cn

Corresponding author: lixun@xpu.edu.cn

In this paper, a method of vehicle multi-object identification and classification based on the YOLOv2 algorithm is proposed, which is used to solve the classical multi-object classification problems of low detection rate, poor robustness and unsatisfactory effect on real road environment. We analyzed vehicle objective and training results. The network structure of YOLOv2-voc is improved according to the actual road conditions based on the YOLOv2 algorithm, and the classification training model was obtained by the ImageNet data which is came from many tweaks. A classification network structure YOLOv2-voc_mul is obtained for sensitive vehicle type changing. In order to verify the validity of the detection method, experiments are performed using samples from simple backgrounds and complex backgrounds and compared with the existing YOLOv2, YOLOv2-voc, YOLOv2-tiny, YOLOv3 and YOLOv3-tiny models after 70000 iterations, respectively. The results show that the proposed YOLOv2-voc_mul model has an accuracy of 98.6% under the simple background, and the mAP (mean Average Precision) of different models reaches 87.81%. Under the complex background, the improved YOLOv2-voc_mul model has an average accuracy of 92.09% and 89.64% for single and multi-object detection of four different models.

KEYWORDS: Intelligent traffic; Multi-object recognition; Convolutional neural network; YOLOv2; Deep learning.

1. Introduction

Image detection and recognition have been an important research in the field of computer vision and machine learning [28]. Detection results of objects are affected by several factors in real scenarios [20], such as: illumination, angle, deformation, occlusion, etc. [31]. Since recognition depends on the results from detection generally, the ability of detection object is vital. Vehicle detection is one of the essential but challenging tasks for traffic and emergency monitoring. In the intelligent transportation system, people can collect traffic information, and the traffic information are acquired from the monitoring system in real time. The monitoring system is used for traffic flow systematic management generally on the roads, such as traffic flow analysis system, license plate detection, highway toll system according to vehicle types and behavior detection of illegal traffic, etc. [1]. This field has been studied actively over the past decades. Many object detection algorithms have been proposed. Among them, the Viola and Jones (V&J) scheme used a sliding-window search with a cascade classifier to achieve accurate location and efficient classification [26], which combines Adaboost with Haar-like features [3]. It has realized real-time face detection and applied in many fields wildly. However, only when it knows the direction of the object can it detect moving objects. Because the original V&J scheme is sensitive to object orientations. Also, a simple decision tree is adopted in the algorithm, which is prone to over-fitting, and it is not ideal for the treatment of complex cases. A linear support vector machine (SVM) classifier with histogram of oriented gradients (HOG) features is used in the other algorithm [24], which has been applied in pedestrian detection successfully [2]. But the detection performance of this method dropped sharply for object with a large intra-class variation. A method named DPM (deformable parts model) had been proposed by P. Felzenszwalb in 2010 for the problem caused by the change of objects appearance during the detection process [6]. DPM is a robust detection method based on object deformation. The idea of improved HOG function, SVM classifier and sliding window detection is adopted by DPM. But this method is necessary to design incentive template manually for different objects [5].

These classical algorithms are divided into three parts mainly. Firstly, the model and extract candidate regions from the scene is established. Then, the feature extraction is carried out for candidate regions, which have been identified. Finally, the objects are classified and the location of valid candidate regions are optimized. In addition, these classical algorithms have some common flaws. The region selection strategy is not targeted based on sliding window, the time complexity is high and the window is redundant. The characteristics of manual design are not robust obviously to objects with diverse variations. Also, these methods are easily influenced by external factors. The method of optical flow for vehicles moving object detection is easy affected by light [15]. Since the changing light will be identified as optical flow mistakenly, this method is not suitable for situations that require harsh real-time performance. It will affect the recognition effect. The frame difference method is easy affected by external noises and speed of detection object [4]. For the objects with different speeds, the detection effect of this method depends on the selected time interval between frames. If the speed of object is fast and the time interval is set larger, it will be detected as two separate objects. The method of background subtraction such as Zhou et al. [30] used the rough-level features to discard the error objects , used the refined-levels features to object matching and used Gaussian functions to perform multi-level feature extraction for discontinuous objects in remote sensing images. Although this method has improved the detection accuracy, large background changes will greatly affect the detection and tracking results.

After several years of research, experiments show that deep convolutional networks can learn very robust and expressive feature representations [14]. Girshick et al. [8] applied deep convolutional network to object detection and achieved great success in the 2012 ImageNet classification task [13]. Then the object detection algorithm of R-CNN (regional convolutional neural network) had been proposed for the first time in 2014. Since then, deep learning has been applied to object detection increasingly such as face recognition [16]. The object detection methods based on deep learning replaced the classical methods gradually. With the increase of data, the performance of classical detection methods tends to bottleneck and the performance of these methods significantly cannot be improved by data augmentation. On the contrary, the performance of deep learning detection methods will be better and better.

In this paper, we propose a new algorithm, which can enhance the detection and recognition accuracy for multiple objects in actual road environment. The purpose of the improved algorithm is to enhance the accuracy of vehicle identification by modifying the network structure many times. We compared the accuracy of the proposed algorithm with other models, such as YOLOv2, YOLOv2-voc YOLOv2-tiny, YOLOv3 and YOLOv3-tiny under simple background and complex background.

The research content of this paper was organized as follows: some related works are studied in the second section. In the third sections, the basic concepts of the YOLO algorithm, the operating rules, and the proffered improved model YOLOv2-voc_mul were discussed thoroughly. In the section 4, experiments and results were shown. At the end of the paper in Section 5, the conclusion of this research was presented.

2. Related Work

In recent years, deep learning methods have become the mainstream object detection method gradually, and several experts take advantage of these methods to detect vehicles [29]. In addition, deep learning methods have also played an important role in autonomous driving and objects tracking. F. Shi has taken advantage of R-CNN for the detection of vehicle and pedestrian in city [25]. But the algorithm takes a long time and the detection speed is slow, which is not suitable for real-time detection in modern traffic. Then K. He et al. [10] proposed the SPPNet and it solved the problem well. However, it still uses the traditional training method and has the large amount of computation, the regression problem about bounding box and classification is performed separately. In 2015, Girshick et al. [9] proposed the Fast-RCNN detector based on R-CNN and SPPNet (spatial pyramid pooling network), which combines the advantages of R-CNN and SPPNet successfully. Then S. Ren et al. [22] proposed Faster-RCNN algorithm, which is the first end-to-end real-time deep learning object detection algorithm. López-Sastre et al. [19] proposed a method to solve the multi-vehicle detection and tracking problem in traffic monitoring applications, which combines SVM and HOG detectors with Faster R-CNN deep learning model. But it takes much time to detect the objects. To shorten the detection time, Joseph et al. [21] proposed YOLO (You Only Look Once) algorithm, which is the first integrated convolution network detection algorithm in 2015. After that, more algorithms were proposed, such as YOLOv2, YOLOv3, SSD (Single Shot Multibox Detector) and etc. [12, 17, 18]. Kim J et al. [11] proposed a method that using multiple sensors to estimate vehicle position during autonomous driving for detecting and tracking moving objects in 2019. Although this method ignores the classification accuracy, the real-time performance is better. The deep convolutional neural networks have a certain degree of invariance to deformation, illumination and geometric transformation. [7, 27]. It can overcome the difficulties effectively brought by the change of vehicle appearance to the object detection and identification. Among them, YOLO algorithm has outstanding performance, simple network structure, faster detection speed, and it can meet the requirements of video detection.

In summary, the YOLOv2 object detection algorithm is used to solve the low recognition rate problem, which caused by vehicle shape, structure color and real scene complex features [23]. According to the characteristics of the vehicle found in the experiments: we fine-tuned the parameters of the network model several times and trained again, and obtained an improved model, which suitable for multi-object vehicle type real-time recognition. In addition, we carried out the experiments of the proposed method in the actual traffic environment, and analyzed the data. The experiment results show that the model has applicability and advancement to the actual traffic environment by comparing with YOLOv2, YOLOv2-voc, YOLOv2-tiny, YOLOv3 and YOLOv3-tiny models.

3. Improved Model of YOLOv2 Based on Empirical Data

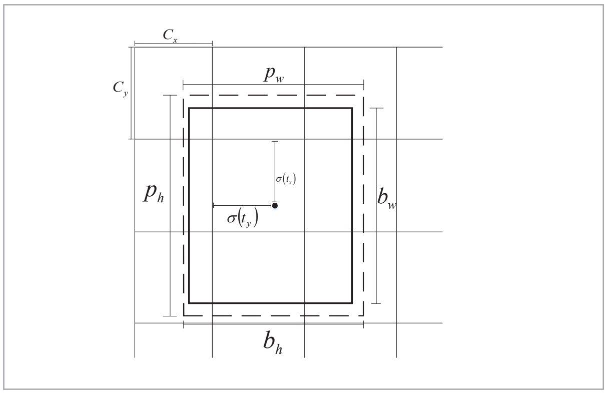
YOLO is an end-to-end object detection network that can detect the object in real time, but for detection performance, YOLOv2 is ore stable than YOLO. R-CNN and Fast R-CNN use selective search to generate candidate frames, and Faster R-CNN uses region candidates to extract candidate frames. YOLO algorithm uses the regression method to return the position and category of the candidate box in the output layer. YOLO algorithm is improved for the detection speed by the regression method, but detection accuracy is lower. YOLOv2 can predict an independent category for each candidate frame and improves the network's ability to detect multiple objects. Therefore, the accuracy of object detection is improved based on maintaining the original high detection speed. YOLOv3 has better detection accuracy and recognition effect when it is detecting tiny objects at a long distance. But the effect of YOLOv3 is lower than YOLOv2 when detecting large objects at close distances.

3.1. Object Detection Algorithm Based on YOLOv2 Model

YOLOv2 is a real-time object detection algorithm, which inputs a picture and outputs the position of the object and the confidence score of the position directly. The sliding window for feature extraction is not used and the classifier is removed the in YOLOv2. The input images are divided into several regions by the algorithm. If the center of a label object falls on a certain area, the object be predicted by this area. The position and confidence of the bounding box are predicted and five predicted values are obtained for each bounding box: (t_x, t_y), (t_w, t_k) and *confidence*, as shown in Figure 1. b_w, b_k are defined as the actual height and width of the anchor, σ is the activation function sigmoid, $\sigma(t_x)$, $\sigma(t_y)$ are the offset from the center of each bounding box to boundary, x and $y$are the offset ratio of the bounding box center to the corresponding grid; t_w and t_y are the true width and height relative to the scale of the entire image, w and h are the ratio between the bounding box and the size of the entire image; the distance between the upper left corner of the grid distance image is (c_x, c_y), and length and width of the bounding box is (p_w, p_h) in each corresponding area. The true position of the bounding boxes is shown in Equation (1).

Figure 1

Bounding boxes with size and location prediction



$$\begin{cases} b_x = \sigma(t_x) + c_x \\ b_y = \sigma(t_y) + c_y \\ b_w = p_w e^{t_w} \\ b_h = p_h e^{t_h} \end{cases}, \qquad (1)$$

where *Confidence* is defined as the accuracy of the predicted position of the bounding box through the product of probability and IOU, it is shown as Equation (2):

$$Confidence = Pr(\text{Object}) \cdot IOU_{\text{pred}}^{\text{truth}}, \qquad (2)$$

where $Pr(\text{Object})$ is the probability value of the object in the grid. If there is an object in a grid, the value of $Pr(\text{Object})$ is 1, or the value is 0 and *confidence* is 0. $IOU_{\text{pred}}^{\text{truth}}$ is defined as the ratio of the predicted object frame to the real object frame. $area(\text{box}_{\text{pred}} \cap \text{box}_{\text{truth}})$ is the area of the intersection of the prediction object box and the real object frame. $(\text{box}_{\text{pred}} \cup \text{box}_{\text{truth}})$ is combined area of the prediction object box and the real object frame as shown in Equation (3):

$$IOU_{\text{pred}}^{\text{truth}} = \frac{area(\text{box}_{\text{pred}} \cap \text{box}_{\text{truth}})}{area(\text{box}_{\text{pred}} \cup \text{box}_{\text{truth}})}, \qquad (3)$$

when the object falls in the grid, the object category is predicted and it is expressed with conditional probability $Pr(\text{class} \mid \text{object})$. The confidence $C(M)$ of certain category M can be obtained through multiplying the predicted probability of class by the category with *Confidence* of the candidate box as shown in Equation (4):

$$\begin{aligned} &Confidence(\mathrm{M}) \\ &= Pr(\text{class} \mid \text{object}) \cdot Pr(\text{object}) \cdot IOU_{\text{pred}}^{\text{truth}} \\ &= Pr(\text{class}_{\mathrm{M}}) \cdot IOU_{\text{pred}}^{\text{turth}} \end{aligned} \quad (4)$$

3.2. Improvement of Multi-object Detection and Recognition Model Based on YOLOv2-voc

In the YOLOv2 algorithm, the anchor box is used to predict the bounding boxes (bboxes) and the last fully connected layer is deleted. The network structure is composed of a convolutional layer and a pooling layer. The picture size is adjusted from 448*448 to 416*416. The image is performed 32 times down sampling and the final output feature size is 13*13. There is a center grid, which is used to predict objects that fall in the center of the image. We adjusted the parameters in the network and conducted several experiments based on the network structure of YOLOv2-voc to get a network with better multi-object detection and recognition as shown in Table 1. The different test results are shown in Figure 2. (the green box is defined as van and the blue box is defined as car.)

In Table 1, model 1 is the YOLOv2-voc, which is consisted of 5 maximum pooling layers and 23 convolutional layers. The Linear activation function and the initial learning rate is 0.001; the initial learning rate is adjusted to 0.0001 in var-model 1 and the initial learning rate is changed to 0.01 in var-model 2 on the basis of model 1; in var-model 3, an average pooling layer is added on the basis of model 1, and 3 convolution layers are removed; in var-model 4, we removed three convolution layers based on the model 1; in var-model 5, we changed the last layer activation function to Rule on the basis model 5; in var-model 6 and var-model 7, we adjusted the number of convolution layers to 21 and 22, and the number of BN layers to 20 and 21, respectively.

From the test results as shown in Figure 2, objects are not detected in subgraph (b); there is a seriously missed detection in subgraph (c), (d) and (f); in subgraph (a), (e), (g) and (h), there are also missed detection when detecting tiny objects in the distance, but the effect of subgraph (e) detection is better among them obviously. In short, we can know that modifying the number of convolutional layers, the average pooling layers, the maximum pooling layers, the BN (Batch Normalization) layers and adjusting the activation function in the network will cause serious missed detection.

If the initial learning rate is set to a small value, there will be repeated detection. If the learning rate is too large, the characteristics of the object cannot be learned, and there will be no detection effect. If the initial learning rate is set to a small value, there will

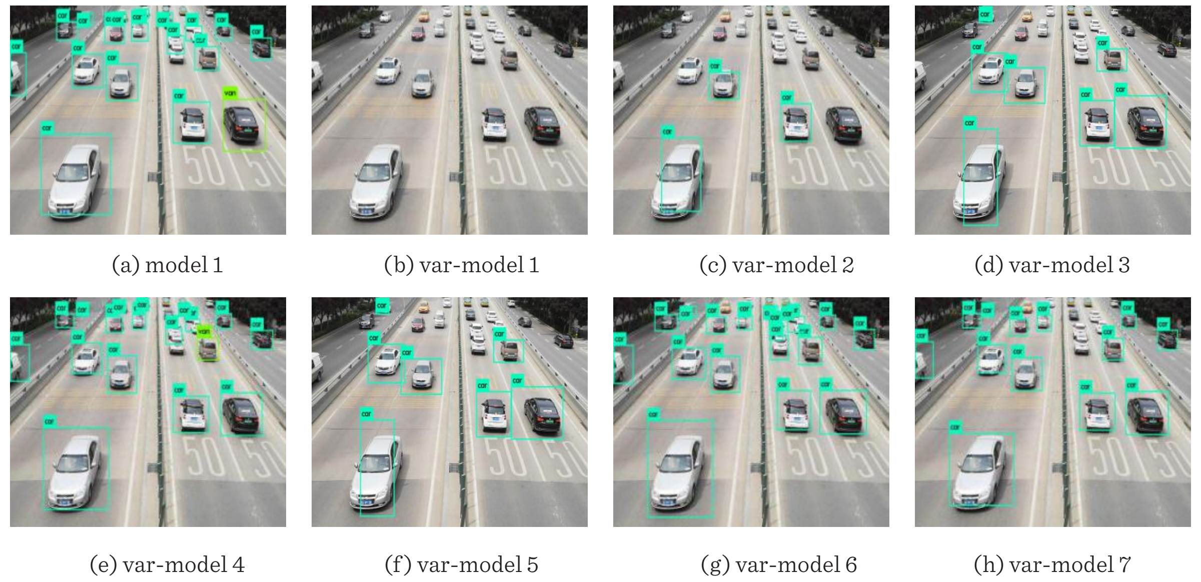
Table 1

Different Network framework

Mode Label	Network Structure				
	Learning Rate	Convolution layer	Maximum Pooling layer + Average Pooling layer	BN layer	Activation Function
model 1	0.001	23	5+0	22	22Leaky+1Linear
var-model 1	0.0001	23	5+0	22	22Leaky+1Linear
var-model 2	0.01	23	5+0	22	22Leaky+1Linear
var-model 3	0.001	20	5+1	19	19Leaky+1Linear
var-model 4	0.001	20	5+0	19	19Leaky+1Linear
var-model 5	0.001	20	5+0	19	19Leaky+1Relu
var-model 6	0.001	21	5+0	20	20Leaky+1Linear
var-model 7	0.001	22	5+0	21	21Leaky+1Linear

Figure 2

The results of different network structure



(a) model 1

(b) var-model 1

(c) var-model 2

(d) var-model 3

(e) var-model 4

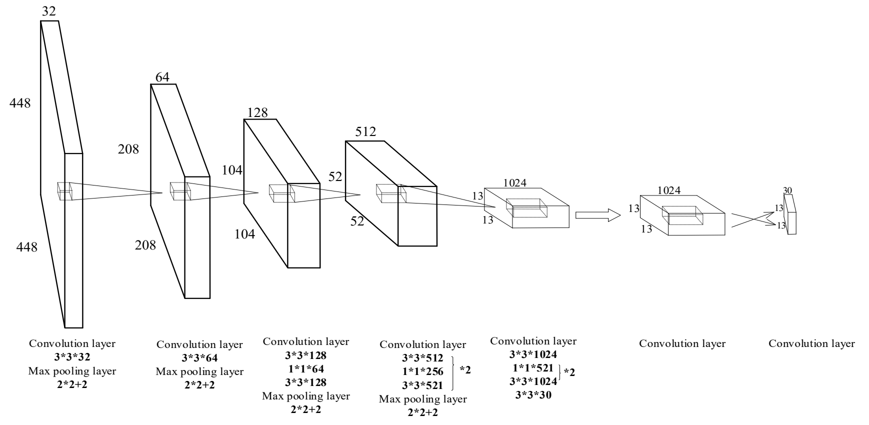
(f) var-model 5

(g) var-model 6

(h) var-model 7

Figure 3

YOLOv2-voc_mul model frame



be repeated detection. If the learning rate is too large, the characteristics of the object cannot be learned, and there will be no detection effect.

To verify the validity of the model, we enhanced the data of sample before training and enriched the data diversity. In deep convolutional neural networks, the purpose of the convolutional layer is to extract deeper features. The deeper the convolutional layer, the smaller the features will be extracted. As the number of convolutional layers increases, the number of neurons increases and the parameters in the network increase. It increases the complexity of the model, causes more difficult to adjust parameters, and the result is more prone to over-fitting. Therefore, we removed the convolutional layers of layers 19, 20, and 21 to get the network structure, which can decreases the complexity of network and reduce the amount of calculation according to the network structure of YOLOv2-voc, as shown in Figure 3. We obtained an improved network model suitable for multi-object vehicle identification and named this network model as YOLOv2-voc_mul (YOLOv2-voc_multiple object vehicle model detection). This network is composed of 20 convolution layers, 5 maximum pooling layers and 19Batch Normalization layers.

4. Experimental Results and Analysis

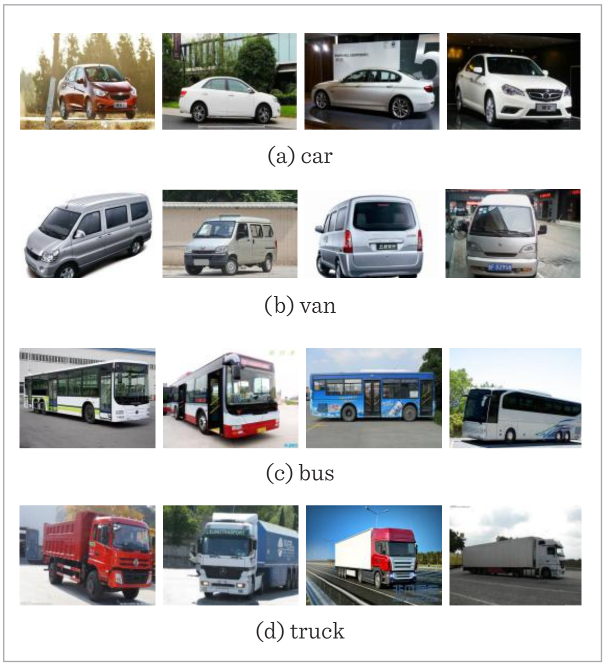
In the process of experiment, a workstation is equipped with an Intel i7-6800 CPU, one NVIDIA GeForce Titan X 1080TI 11GB GPU and eight 32GB memories.

4.1. Data Sample

Vehicle characteristics need be learned from a large number of samples in vehicle object recognition method based on Convolutional Neural Network. If the sample is not represented, it is difficult to select good features. To verify the effectiveness of the proposed method, we collected mixed vehicle data from a variety of car websites. The VOC data set is composed of four different types of vehicles (i.e., car, van, bus, truck), and 2000 samples for each type, and we calibrated the locations of all data samples. To meet the requirements of the basic data volume, we have extended the sample size to five times as our training data set by data enhancement. Some data samples are shown in Figure 4.

Figure 4

Data sample for train from Internet



(a) car

(b) van

(c) bus

(d) truck

4.2. Analysis of Experimental Results

4.2.1. Analysis of Model Verification Results

In this part, we compared the typical model YOLOv2, YOLOv2-voc, YOLOv2-tiny, YOLOv3 and YOLOv3-tiny with the proposed YOLOv2-voc_mul model. We initialized the parameters of these models by using the pre-trained network model. During the training, two image samples are input each time, weight is updated after once per iteration, and every ten iterations is a cycle. We set the initial learning rate to 0.001 and change 0.01 to 0.0001, 0.00001 and 0.000001 in the 10000, 20000 and 40000 iterations, respectively, to obtain multi-object system detection models with different weights. Among them, there is "Nan" during the training process in YOLOv3 model, because the sample vehicle in the data set are large objects in a simple background. The model is more suitable for the detection of tiny objects. Finally, we obtained the multi-object detection effect by comparing with the verification set.

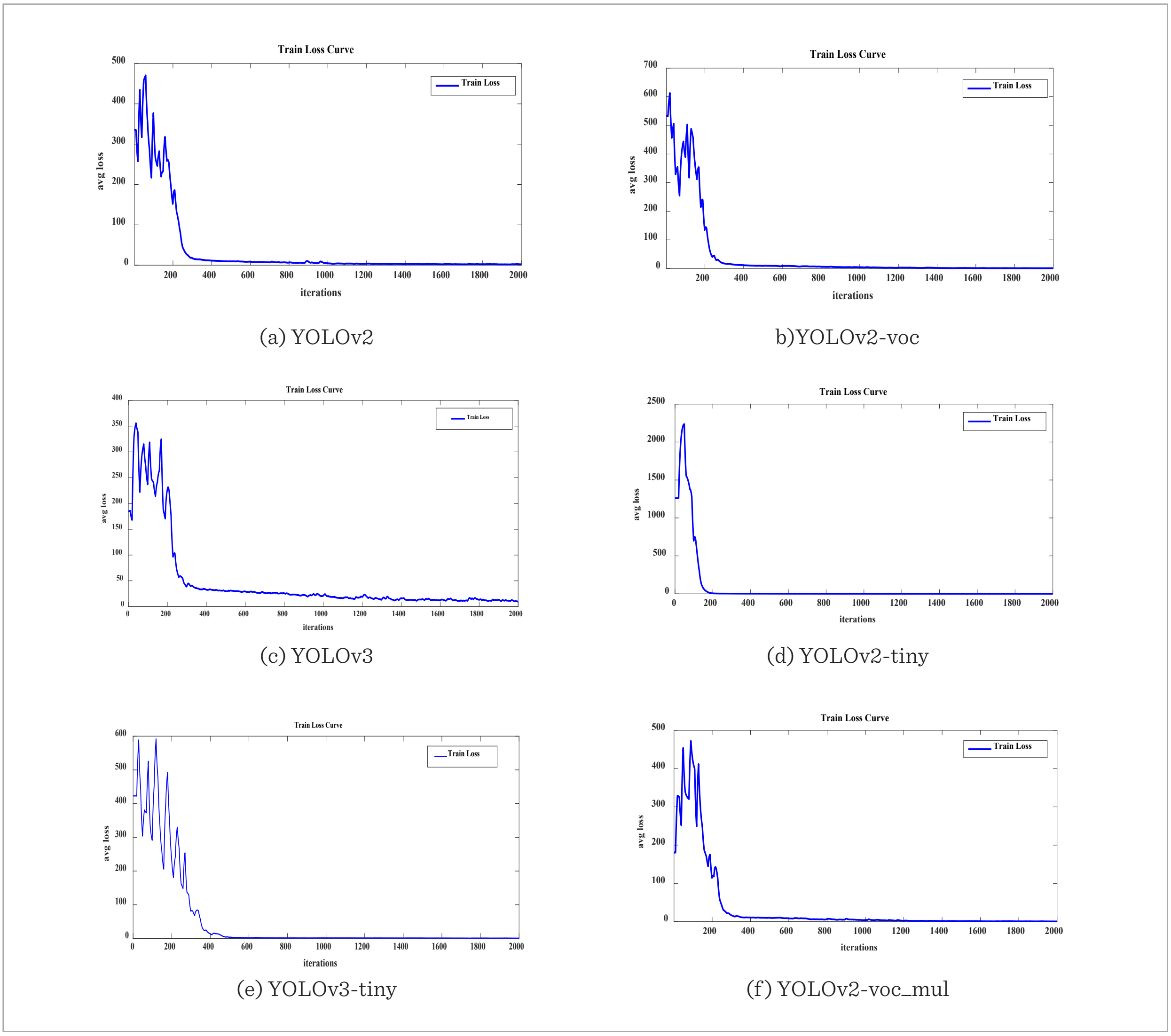
(1) Loss curve analysis

Figure 5 are loss graphs of YOLOv2, YOLOv2-voc, YOLOv2-tiny, YOLOv3, YOLOv3-tiny and YOLOv2-voc_mul models during training.

In Figure 5, the subgraph (a), (b), (c), (d), (e) and (f) are the loss graphs of the YOLOv2, YOLOv2-voc, YOLOv2-tiny, YOLOv3, YOLOv3-tiny and YOLOv2-voc_mul models respectively. All the training times are 2000 at an initial learning rate. The divergence is more serious due to the low learning at the beginning of training. The maximum loss value of YOLOv3 has reached more than 2000 at the beginning of the training, which is the highest among all models. The maximum loss values of other models are between 400 and 600. The values are in a normal range. In the aspect of convergence speed, YOLOv3 converges to approach 0 after about 200 iterations. YOLOv2, YOLOv2-voc and YOLOv2-voc_mul models start to converge in about 350 iterations and approach to 0 after 1200 iterations. After about 400 iterations, YOLOv2-tiny model converges to 0 and YOLOv3-tiny model begins converge at around 400 iterations. From the Figure 5, the loss curves of the five models have only a slight difference

Figure 5

Loss graphs of six different models after 2000 iterations



(a) YOLOv2

b)YOLOv2-voc

(c) YOLOv3

(d) YOLOv2-tiny

(e) YOLOv3-tiny

(f) YOLOv2-voc_mul

except YOLOv3 model. During the progress of experiments we found that the difference convergence speed in the early has little effect on the recognition effect in the later stage. Therefore, YOLOv3 model has no absolute advantages for the vehicle identification.

(2) Accuracy analysis

In Table 2, *Total* is the actual number of objects to be detected; *Correct* is the number of detected bboxes by the network after importing the picture. Each bbox has corresponding confidence. When the confidence is bigger than the threshold, the *IOU* need be computed to find the bbox which has the largest *IOU*. If the bbox of maximal *IOU* is bigger than the set *IOU* threshold, the *Correct* value will increase by 1. *Proposal* means the number of bboxes which is bigger than the threshold in all detected bboxes; *Precision* is the accuracy of the model as defined in Equation (5); *Recall* is defined as the ratio between the number of detection objects and the number of all objects as shown in Equation (6); F_1 means the balanced F Score. It is defined as the harmonic average between the accuracy rate and the recall rate and the recall rate and accuracy of the model are taken into account. The range of value is from 0 to 1. The higher F_1, the better effect, as shown in Equation (7).

$$Precision = \frac{Correct}{Proposal}, \tag{5}$$

$$Recall = \frac{Correct}{Total}, \tag{6}$$

$$F_1 = 2 \cdot \frac{Precision \times Recall}{Precision + Recall}. \tag{7}$$

It can be seen from Table 2: when verifying total objects, 84 objects can be detected accurately by YOLOv3 model. The initial learning rate of the model is set to 0.001, but the *Precision* is 55.63%, the *Recall* is 54.55%, and the F_1 is 55.08%. The three indicators of YOLOv2-tiny model have not good performance. The *Precision* of YOLOv3-tiny model is better than others. The results show that the detection effect of the three models is not good. 152 objects can be detected accurately in the YOLOv2 model, the *Precision* reaches 96.71%, *Recall* reaches 95.45% and the F_1 is 96.07%; the *Precision* of YOLOv2-voc model is improved to 97.28%, but *Recall* and F_1 have decreased slightly. The test results of the improved model YOLOv2-voc_mul show that *Precision* has been increased to 98.62%, the *Recall* has been increased to 94.81% compared with YOLO v2-voc and the F_1 is increased to 96.67%. The three aspects have been enhanced in different degrees. Compared with the other five models, the better detection results can be obtained. To make the information more intuitive in Table 2, we plot the curve of *IOU*, *Recall* and *Precision* as shown in Figure 6.

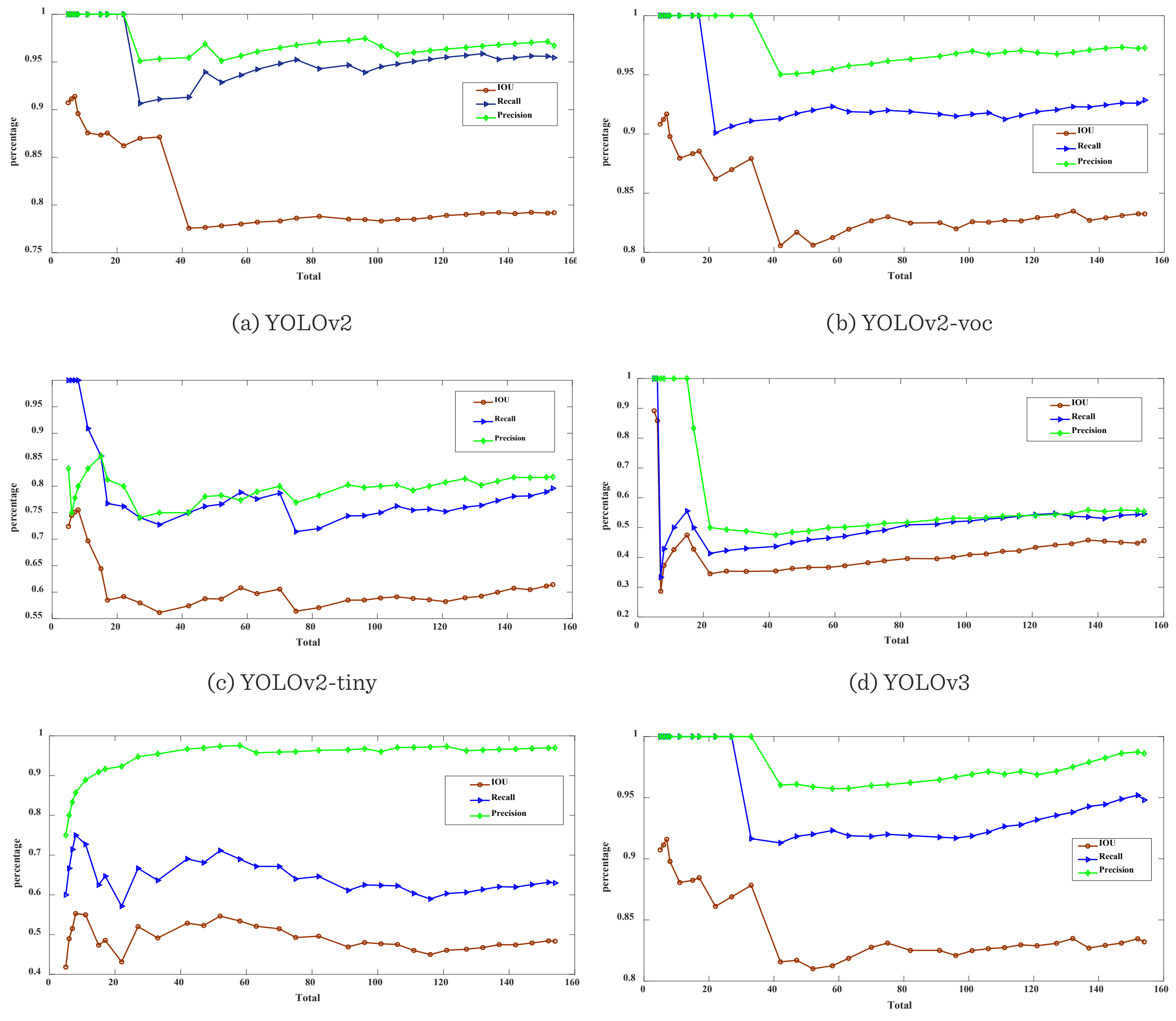
In Figure 6: the recall rate of the six models show fluctuations significantly at the beginning. When the number of detected objects increases, the recall rate of the YOLOv2 model stabilizes at 95.5% gradually, the recall rate of YOLOv2-voc model tends to 93%, the recall rate of YOLOv2-voc_mul model is stable at 94.8%, the recall rate of the YOLOv2-tiny model tends to 80%, the recall rate of the YOLOv3-tiny model tends to 63%, and the recall rate of the YOLOv3 model is fluctuating between 40% to 45%. It proves that the better correct rate can be guaranteed in YOLOv2, YOLOv2-voc and YOLOv2-voc_mul models, but the correct rate of YOLOv2-tiny, YOLOv3 and YOLOv3-tiny models are bad. In the aspect of precision, the curves of the YOLOv2 and YOLOv2-voc model have a significant fluctuation when the number of objects increases, the precision of YOLOv3-tiny model is tend to 97%. The precision of improved YOLOv2-voc_mul model is stable at around 98.6% after relatively small fluctuations and good accuracy and stability are kept during training. However, the precision of YOLOv2-tiny, YOLOv3 models have dropped dramatically. Comparing the *IOU* curves of the six models, the *IOU* of YOLOv2 model fluctuates around 0.80, the *IOU* of the YOLOv2-voc model and the YOLOv2-voc_mul model are improved compared to YOLOv2, and it maintained at around 0.83. But the *IOU* of YOLOv2-tiny model fluctuates around 0.61, the *IOU* of YOLOv3 and YOLOv3-tiny models fluctuate between 0.4 and 0.5. They are the worse stability compared to the other three models.

Table 2
Test results of different models

Model	Total	Correct	Proposal	Precision(%)	Recall(%)	F_1 (%)
YOLOv2	154	147	152	96.71	95.45	96.07
YOLOv2-voc	154	143	147	97.28	92.86	95.01
YOLOv2-tiny	154	121	148	81.76	79.61	80.63
YOLOv3	154	84	151	55.63	54.55	55.08
YOLOv3-tiny	154	97	100	97.00	62.99	76.38
YOLOv2-voc_mul	154	146	148	98.62	94.81	96.67

Figure 6

The verification results of different models



(a) YOLOv2

(b) YOLOv2-voc

(c) YOLOv2-tiny

(d) YOLOv3

(e) YOLOv3-tiny

(f) YOLOv2-voc_mul

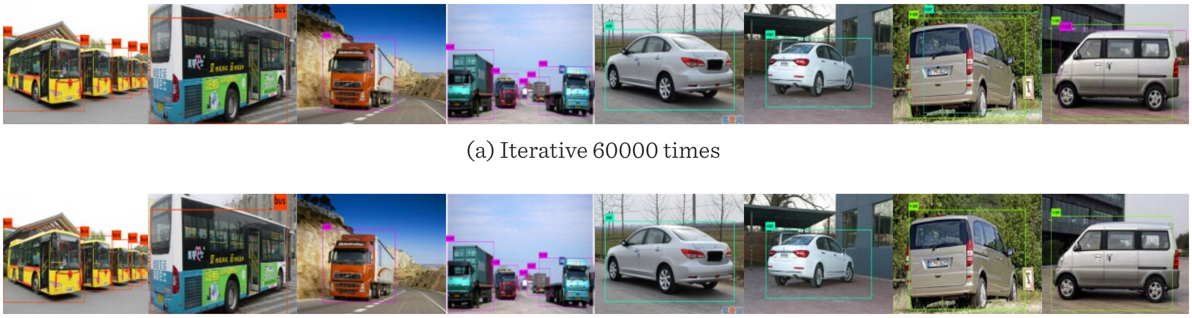
(3) Analysis of test results

In order to verify the effect preferably, the YOLOv2-voc_mul model is trained to obtain the weight of 60,000 times and 70,000 times after training. We used different weights to perform vehicle multi-target detection and analyzed the test results.

From the Figure 7: the red box is defined as bus, the pink box is defined as truck, the blue box is defined as car and the green box is defined as van. When YOLOv2-voc_mul is trained for 60,000 times, the van is recognized as truck or car. This is a phenomenon of obvious false detection. Because the feature of van is not obvious enough such as small shape, appearance features are not prominent. It has similarities features with car and truck. After 60,000 iterations, the more detailed features of the objects are not learned by the model. The false detection phenomenon is eliminated after the number of training 70,000 times. We can see that the YOLOv2-voc_mul model can learn more comprehensive features, reduces false detection, identifies four different models of vehicles

Figure 7

The results of different iterative times



(a) Iterative 60000 times

(b) Iterative 60000 times

correctly (i.e., truck, bus, van, and car) after training 70,000 times, the effect of recognition is better too.

4.2.2. Result of Experimental Results and Analysis

In the experiments of vehicle type identification detection, we compared the AP values of YOLOv2, YOLOv2-voc, YOLOv2-tiny, YOLOv3, YOLOv3-tiny and YOLOv2-voc_mul models and analyzed the data. After 70,000 iterations of the six models, we got the initial recognition detection models, and the detection results are shown in Table 3. And the Figure 8 is the AP value graph of the van.

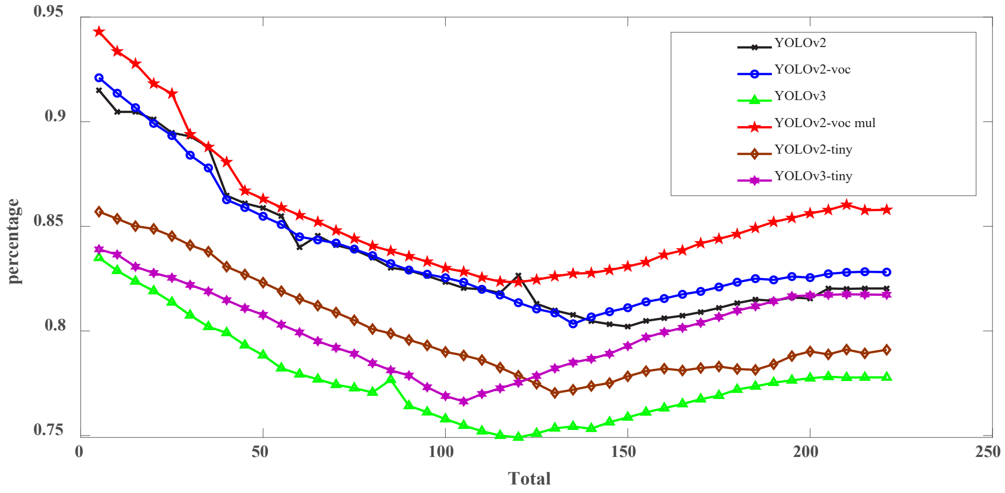
Table 3

The results of different models detecting AP values (%)

model	truck	bus	van	car	mAP
YOLOv2	85.37	83.93	82.03	86.05	84.35
YOLOv2-voc	86.83	84.74	82.81	86.51	85.22
YOLOv2-tiny	83.21	82.88	79.10	84.22	82.35
YOLOv3	82.44	81.23	77.80	82.16	80.91
YOLOv3-tiny	83.44	82.03	78.13	83.38	81.75
YOLOv2-voc_mul	87.91	87.76	85.79	89.80	87.84

Figure 8

The AP value graph of van in different models



From the Table 3, the results show that the mAP of YOLOv2 is 84.35% for the four vehicle types, the mAP of YOLOv2-voc has been increased by 0.87% and the mAP of YOLOv2-voc_mul has been increased to 87.84%. The other models' mAP are relatively lower. The van is more challenging to distinguish. The AP of YOLOv2 is 82.03%, the AP of YOLOv2-voc is 82.81%, the AP of YOLOv2-tiny, YOLOv3 and YOLOv3-tiny are less than 80%. The model we proposed has been increased to 85.79%. Therefore, the YOLOv2-voc_mul model has a higher recognition rate and better classification effect for different vehicle types obviously. In Figure 8, we can see that the graph of YOLOv2-voc_mul is the best one and the curves of YOLOv2-tiny, YOLOv3, YOLOv3-tiny are worse. The curve of YOLOv2 is similar to the curve of YOLOv2-voc.

4.3. Multi-object Recognition Test Results and Analysis

The YOLO algorithm is affected by the training samples greatly, so we need the rich, diverse and representative samples. The single-object features are obvious in the simple background, and these features are learned easily during the training process, so the recognition rate is better. The background of road objects is very complex and there are many disturbances. The process of learning features is difficult relatively and the recognition rate is also affected. Therefore, we expanded the training sample and increased 2000 samples of different vehicle types which taken by cameras on the overpass. As shown in Figure 9.

Figure 9

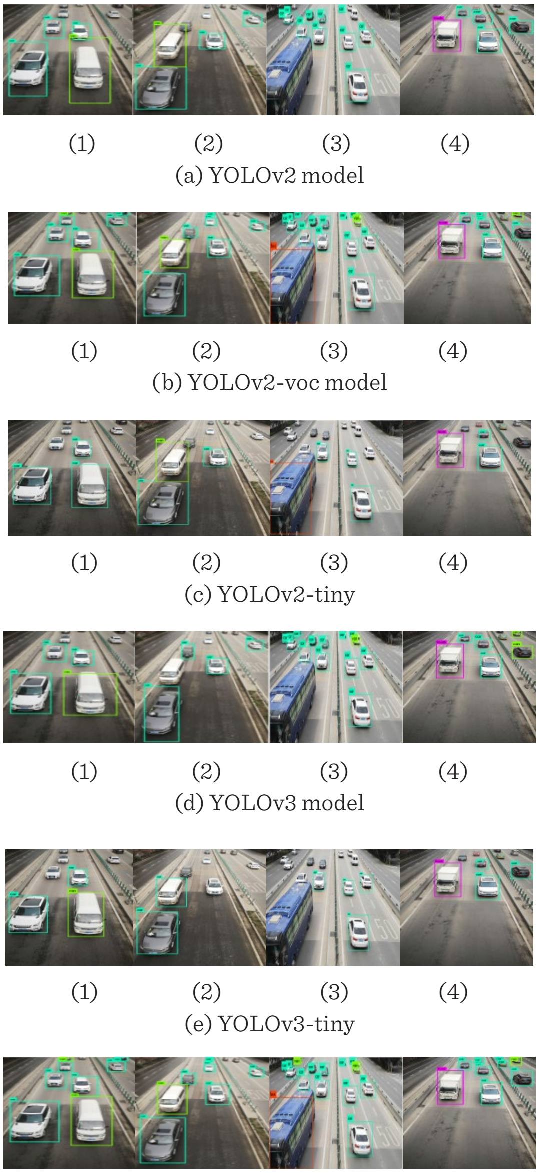
Data sample from actual traffic environment

4.3.1. Test Results of Actual Road Object

We trained the data set for 70000 times after increasing the actual road object samples. Through experimental testing, we obtained the results of the verification set, as shown in Figure 10. And we made the confusion matrix, as shown in Figure 11.

Figure 10

The test results of different models



(1) (2) (3) (4)

(a) YOLOv2 model

(1) (2) (3) (4)

(b) YOLOv2-voc model

(1) (2) (3) (4)

(c) YOLOv2-tiny

(1) (2) (3) (4)

(d) YOLOv3 model

(1) (2) (3) (4)

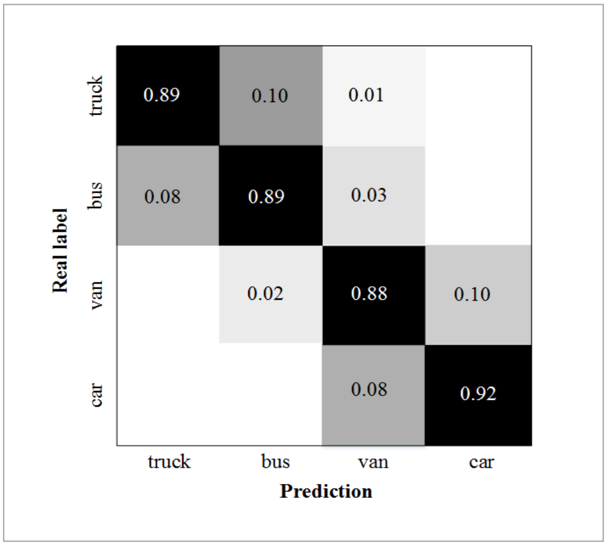
(e) YOLOv3-tiny

(1) (2) (3) (4)

(f) YOLOv2-voc_mul model

From the Figure 10, in subgraph (a), there is a seriously missed detection and the distant tiny objects are not detected in the YOLOv2 model. In subgraph (b), all objects of (1) are detected; there is a false detection in the upper left corner, and the "van" is recognized as "car" in (3). In subgraph (d), there is also a seriously missed detection, and it has a false detection, which recognized "van" as "car". In subgraph (c) and (e), there are false detection about "car" and "van", and there are seriously missed detection. In subgraph (f), it eliminates the miss detection phenomenon and detects the two "van" objects in the distance correctly. It can be seen that our proposed model has a better efficiency in the actual road object detection. In Figure 11, the row is defined as the actual category and the column is defined as the predicted results. From the Figure 11, it is on the diagonal roughly and the classification result is good. Van is the key reason and the accuracy of the car is affected, because their types and sizes are similar. Also, bus, truck and van affect each other.

Figure 11
Confusion matrix



	truck	bus	van	car
truck	0.89	0.10	0.01	
bus	0.08	0.89	0.03	
van		0.02	0.88	0.10
car			0.08	0.92

4.3.2. Analysis of Single Object and Multi-object Detection Results

We used the expanded data set to perform classification experiments for single-object and multiple-object to prove the applicability of the improved model. And we trained the model 70,000 times during the experiment.

Table 4
The average accuracy of vehicle identification (%)

Type	truck	bus	van	car	average accuracy
simple	92.03	91.93	89.88	94.52	92.09
multiple	88.91	88.86	88.07	92.71	89.64

When detecting single and multiple objects, the average accuracy are 92.09% and 89.64%. From the Table 4, the average accuracy of the truck is 92.03%, the average accuracy of bus is 91.93%, the average accuracy of van is 89.88% and the average accuracy of car is 94.52%. In multi-object detection, the average accuracy of truck is 88.91%, the average accuracy of bus is 88.86%. The average accuracy of van and car have been decreased by 1.81% compared with the average accuracy of single-object detection. In a word, the accuracy in the multi-object is lower than single-object in the simple background.

5. Conclusion

In this paper, The parameters of YOLOv2, YOLOv2-voc, YOLOv2-tiny, YOLOv3 and YOLOv3-tiny models were analyzed firstly, including learning rate, the number of convolutional layers, batch normalization layers, pooling layers, and activation function. The same set of samples have been compared in different models, and detection results were compared, according to object's appearance characteristics and their movement characteristics. Through a lot of experiments in different backgrounds, the detection accuracy of the proposed method can reach 98.62% in the simple background and 89.64% in a complex background for single object. The accuracy of the van that is difficult to identify is increased to 88.96%. The detection accuracy and operation efficiency are improved in the proposed method. Compared with YOLOv2 YOLOv2-voc, YOLOv2-tiny, YOLOv3 and YOLOv3tiny models, the results show that the detection and recognition effect of this method has better performance used for identification of vehicle types on the actual road.

6. Discussion

In this paper, the model framework and the amount of calculation of parameters are reduced, and the accuracy is improved. But the proposed method still has limitations: Although we can improve the accuracy, the detection of distant tiny objects are not accurate enough. When we use this method to detect objects, there may still be missed detection of distant tiny objects. For the reasons stated above, we would focus on tiny objects detection and improve the accuracy of results. Vehicle types classification intensively also a question worth studying. In the future, we will combine this classification method with multi-object tracking methods to achieve real-time tracking of moving objects.

Acknowledgement

This study was supported by the national natural science foundation of China (No.61971339 and No.51905405); the basic research program of natural science of Shaanxi province (No.2019JM567) and the guiding program of science and technology of China textile industry federation (No.2018094).

References

1. Azam, S., Rafique, A., Jeon, M. Vehicle Pose Detection Using Region Based Convolutional Neural Network. International Conference on Control Automation and Information Sciences, 2016, 194-198.https://doi.org/10.1109/ICCAIS.2016.7822459
2. Bilal, M., Khan, A., Karim, Khan, M. U., Kyung, C. A Low-Complexity Pedestrian Detection Framework for Smart Video Surveillance Systems. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(10), 2260-2273.https://doi.org/10.1109/TCSVT.2016.2581660
3. Chen, S., Ma, X., Zhang, S. AdaBoost Face Detection Based on Haar-Like Intensity Features and Multi-threshold Features. 2011 International Conference on Multimedia and Signal Processing, 2011, 251-255. https://doi.org/10.1109/CMSP.2011.58
4. Fan, J., Gao, Y., Wu, Z., Li, L. Research of Moving Target Detection Method Based on Moving Camera. Electronic Measurement Technology, 2018, 41(01), 129-134.
5. Felzenszwalb, P., Girshick, R., Mcallester, D. Object Detection with Discriminatively Trained Part-Based Models. IEEE Transactions on Software Engineering, 2010, 32(9), 1627-1645. https://doi.org/10.1109/TPAMI.2009.167
6. Felzenszwalb, P., McAllester, D., Ramanan, D. A Discriminatively Trained, Multiscale, Deformable Part Model. 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, 1-8. https://doi.org/10.1109/CVPR.2008.4587597
7. Garcia-Martin, A., Martinez, J. M. Enhanced People Detection Combining Appearance and Motion Information. Electronics Letters, 2013, 49(4), 256-258. https://doi.org/10.1049/el.2012.3817
8. Girshick, R., Donahue, J., Darrell, T. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. IEEE Conference Computer Vision and Pattern Recognition, 2014, 580-587. https://doi.org/10.1109/CVPR.2014.81
9. Girshick, R. Fast R-CNN. 2015 IEEE International Conference on Computer Vision (ICCV), 2015, 1440-1448.https://doi.org/10.1109/ICCV.2015.169
10. He, K., Zhang, X., Ren, S. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9), 1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824
11. Kim, J., Choi, Y., Park, M. W. Multi-sensor-based Detection and Tracking of Moving Objects for Relative Position Estimation in Autonomous Driving Conditions. The Journal of Super computing, 2019. https://doi.org/10.1007/s11227-019-02811-y
12. Kim, K., Kim, P., Chung, Y., Choi, D. Performance Enhancement of YOLOv3 by Adding Prediction Layers with Spatial Pyramid Pooling for Vehicle Detection. 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018, 1-6. https://doi.org/10.1109/AVSS.2018.8639438
13. Krizhevsky, A., Sutskever, I., Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. Communications of the ACM, 2017, 60(6), 84-90. https://doi.org/10.1145/3065386
14. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P. Gradient Based Learning Applied to Document Recognition. Proceedings of the IEEE, 1998, 86(11), 2278-2323. https://doi.org/10.1109/5.726791

15. Li, C., Bai, H., Guo, H., Liang, H. Moving Object Detection and Tracking Based on Improved Optical Flow Method. Chinese Journal of Scientific Instrument, 2018, 39(5), 249-256. https://doi.org/10.19650/j.cnki.cjsi.J1803270

16. Li, X., Li, L., Nan, K. Face Recognition Method Based on Smart Home Mobile Robot. Journal of Xi'an Polytechnic University, 2020, 34(01), 61-66. https://doi.org/10.13338/j.issn.1674-649x.2020.01.010

17. Li, X., Liu, Y., Li, P. F., Zhang, L., Zhao, Z. Vehicle Multi-target Detection Method Based on YOLOv2 Algorithm Under Darknet. Journal Traffic and Transportation Engineering, 2018, 18(06), 142-158. https://doi.org/10.19818/j.cnki.1671-1637.2018.06.015

18. Liu, W., Anguelov, D., Erhan, D. SSD: Single Shot MultiBox Detector. Lecture Note in Computer Science, 2015, 9905, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

19. López-Sastre, R. J., Herranz-Perdiguero, C., Guerrero-Gómez-Olmedo, R., Oñoro-Rubio, D., Maldonado-Bascón, S. Boosting Multi-Vehicle Tracking with a Joint Object Detection and Viewpoint Estimation Sensor. Sensors, 2019, 19(19), 4062. https://doi.org/10.3390/s19194062

20. Murali, S., Govindan, V. K., Kalady, S. A Survey on Shadow Detection Techniques in a Single Image. Information Technology & Control, 2018, 47(1), 75-92. https://doi.org/10.5755/j01.itc.44.1.5757

21. Redmon, J., Divvala, S., Girshick, R., Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 779-788. https://doi.org/10.1109/CVPR.2016.91

22. Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(06), 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

23. Sang, J., Wu, Z., Guo, P., Hu, H. An Improved YOLOv2 for Vehicle Detection. Sensors, 2018, 18(12), 4272. https://doi.org/10.3390/s18124272

24. Saric, M., Dujmic, H., Russo, M. Scene Text Extraction in IHLS Color Space Using Support Vector Machine. Information Technology & Control, 2015, 44(1):20-29. https://doi.org/10.5755/j01.itc.44.1.5757

25. Shi, F. Research on Real-time Identification Method of Pedestrians and Vehicles Based on R-CNN. Harbin Institute of Technology, 2019.

26. Viola, P., Jones, M. Rapid object detection using a boosted cascade of simple features. Proceeding of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001, I-I. https://doi.org/10.1109/CVPR.2001.990517

27. Wang, P., Hao, W., Sun, Z., Wang, S. Regional Detection of Traffic Congestion Using in a Large-Scale Surveillance System via Deep Residual Traffic Net. IEEE Access, 2018, 6, 68910-68919. https://doi.org/10.1109/ACCESS.2018.2879809

28. Xu, Y., Yu, G., Wu, X., Wang, Y., Ma, Y. An Enhanced Viola-Jones Vehicle Detection Method from Unmanned Aerial Vehicles Imagery. IEEE Transactions on Intelligent Transportation Systems, 2017, 18(7), 1845-1856. https://doi.org/10.1109/TITS.2016.2617202

29. Yao, H., Yu, Q., Xing, X., He, F., Ma, J. Deep-learning-based Moving Target Detection for Unmanned Air Vehicles. 2017 36th Chinese Control Conference, 2017, 11459-11463.https://doi.org/10.23919/ChiCC.2017.8029186

30. Zhou, B., Duan, X., Ye, D., Wei, W., Woźniak, M., Połap, D., Damaševičius, R. Multi-Level Features Extraction for Discontinuous Target Tracking in Remote Sensing Image Monitoring. Sensors, 2019, 19(22), 4855. https://doi.org/10.3390/s19224855

31. Zhou, F., Li, J., Li, X., Cao, Y. Freight Car Target Detection in a Complex Background Based on Convolutional Neural Networks. Proceedings of the Institution of Mechanical Engineers, 2019, 233(3), 298-311. https://doi.org/10.1177/0954409718793464