EAI Endorsed Transactions
on Scalable Information Systems

Research Article **EAI.EU**

A novel one-stage object detection network for multi-scene vehicle attribute recognition

Jiefei Zhang[1,*]

[1]School of Automobiles, Henan College of Transportation, Zhengzhou 450000, China

Abstract

In recent years, with the continuous development of computer vision technology, computer vision has been widely used in many scientific research fields and civil applications. As one of the basic tasks of many advanced visual tasks, object detection has important research significance in the field of computer vision and practical applications. At present, with the joint efforts of many scholars, the research on object detection based on deep learning has made remarkable progress. However, in some special weather, such as rainy days, foggy days, nights and the lack of visible light source, the visual distance and visibility are very poor, and the obtained images cannot be used normally, thus affecting the result of object detection. To solve the above problem, this paper proposes a novel one-stage object detection network for multi-scene vehicle attribute recognition, which mainly contains vehicle type and color attributes. The one-stage object detection network YOLOv3 is used as the basic network, and GIOU loss function is used to replace MSE loss function. Finally, experimental results show that the accuracy of the proposed algorithm is improved significantly on public data sets.

Keywords: multi-scene vehicle attribute recognition, YOLOv3, GIOU loss function.

Received on 13 November 2021, accepted on 22 November 2021, published on 22 November 2021

doi: 10.4108/eai.22-11-2021.172217

1. Introduction

With the rapid development of social economy, the living standards of the people have risen steadily. At present, the transportation tools have gradually transformed from their own feet to various motor vehicles, with various kinds of transportation problems becoming more and more complex. Intelligent Transportation Systems (ITS) is an ideal solution specifically to traffic problems caused by current economic development. As an important member of traffic, all kinds of motor vehicles play the effective identification of their attributes in the intelligent transportation system [1-4]. In the general surveillance video screen, there is generally not clear image pixels, the license plate is blocked, daub, corrosion and other situations, which is unable to accurately locate and identify the license plate number, then it is particularly important to quickly and accurately identify the other attribute information of the vehicle. For example, other appearance characteristics of the vehicle, the accurate identification of the vehicle type and color can make up for the lack of license plate identification, and can supplement the license plate identification results, more comprehensively increase the vehicle information. It can also improve the reliability and safety of ITS, and greatly improve the intelligence of vehicle traffic management. Quickly and accurately identifying the type and color of vehicles and making accurate analysis according to the identification results can effectively serve ITS, and can also establish a vehicle information database to provide vehicle information retrieval, which will greatly improve

[*]Corresponding author. Email: aqiufenga@163.com

the work efficiency of relevant departments. Therefore, the effective identification of vehicle attributes plays an extremely important role in traffic congestion, vehicle retrieval and tracking, and real-time detection and management of vehicles, and it will become one of the important auxiliary means of ITS [5,6].

Vehicle color recognition is a sub-field of computer vision research. Its main work is to determine the main color of vehicles in an image. Due to the complex external environment factors such as weather and illumination, traditional color recognition still has many problems, so it can not correctly identify the color of each vehicle. In addition, some vehicles are too close to each other to accurately classify, which is another challenge in color recognition. Vehicle color recognition in natural scenes can provide useful information in vehicle detection, vehicle tracking and automated driving systems. Due to its specific edge features, vehicle types are easier to identify than vehicle colors. As long as there are enough samples and the image quality is high enough, vehicle types can be accurately identified with high accuracy. In the traditional vehicle attribute recognition, it is usually for a single attribute. Zhao et al. [7] proposed a new network based on Capsule Network, which included five different types of vehicles with different angles and illumination conditions. The dynamic routing algorithm of capsule network was redesigned to save training time and speed up convergence. Finally, the accuracy of 91.27% was achieved. Chen et al. [8] divided each vehicle face image into multiple sub-images according to texture features, and used convolutional neural network (CNN) in different layers. CNN extracted global and local features to identify vehicle models, achieving a high accuracy, but it could only recognize the image containing the vehicle foreground, and could not correctly recognize the image from other angles of the vehicle [9-11]. Damitha et al. [12] used the support vector machine (SVM) [13] method to classify and recognize vehicle colors. They extracted the same Region of Interest (ROI) to extract color features from vehicles, several different feature combinations were used, and testings were carried out according to these feature combinations. Finally, 87.52% average accuracy was achieved. Xue et al. [14] used different processing methods for images under different lighting conditions to weaken the influence of lighting factors on color recognition and improve the accuracy of color recognition. However, they needed to spend a lot of time manually classifying images during image processing. Ruan et al. [15] first used faster R-CNN network to detect vehicles, and then improved the structure of GoogLeNet network by connecting multiple loss layers behind the full connection layer to realize identification of vehicle identification, posture and color attributes, which achieved 85% accuracy.

To sum up, there are many studies and achievements in the field of vehicle attribute recognition, but most of the learning tasks are based on single-task learning with single properties, with few studies on complex learning tasks via multiple properties. It is difficult to achieve locating a specific vehicle only through a certain attribute of the vehicle. The identification process generally takes a long time and cannot be effectively applied to practical applications. Moreover, the more information the external attributes of the vehicle can be determined, the greater it helps to locate a specific vehicle. For a vehicle, also has multiple attribute descriptions. For example, according to the type of car, it can be described as car, suv, van, etc. Depending on the color of the vehicle, it can be described as red, white, black, etc. Based on the above analysis, this paper constructs the vehicle multi-attribute data set through network search and real shot data, and proposes a vehicle attribute identification method based on the deep neural network. The upgraded YOLOv3 network is used to train the image global area, the vehicle type and color to detect the vehicle type and color attributes to improve the applicability of the model.

2. YOLOv3 network

At present, deep learning (DL) is widely used in the field of target detection and recognition [16-18]. In recent years, the algorithm based on DL has made breakthrough achievements in image classification and object detection. When processing visual tasks, deep learning architectures can learn more efficient representations from raw images than traditional methods using artificial features, and perform better than traditional methods. CNN first shows its practicality in the digital recognition work. Krizhevsky et al. [19] first applied CNN to large-scale image classification problems, and obtained a high performance application instance on ImageNet dataset, which was the largest and most challenging image dataset so far.

In this paper, the YOLOv3 Network is used as the prototype, and the residual network (ResNet) [20,21] is used to skip the layer connection mode to increase the network depth and still make the network convergence, which achieves end-to-end target detection and identification. YOLOv3 neural network can predict the object and its position information in the image only by looking at the image once. The detection frame is extracted directly from the image and the target object is detected by the whole image feature. In this process, the input image is first divided into S×S grids, and then B detected boxes are predicted for each grid. Each detection box contains five predicted values, namely, x, y, w, h and confidence. x and y are the central coordinates of the detection enclosure. e and h indicate the width and height of the detection enclosure respectively. Confidence is the confidence of the category to which this detection box belongs. The loss function of each detection frame contains four parts, as shown in equation (1):

$$Loss = \lambda_{coord} loss_{xy} + \lambda_{coord} loss_{wh} + loss_{conf} + loss_{cls}$$

(1)

Where $loss_{xy}$ is the center coordinate error of the detection frame. $loss_{wh}$ is the height coordinate error of the detection frame. $loss_{conf}$ is the confidence error of the detection frame. $loss_{cls}$ is the classification error of a detection frame. The loss function is divided into two parts: the part with object and the part without object [22]. The loss of the part without object increases the weight coefficient. Most of the content in an image does not contain the object to be detected, which will lead to the calculation amount of the part without object is greater than that of the part with object, which will lead to the tendency of the network to detect the cell grid without object. Therefore, the weight coefficient is added in the part without objects to reduce the contribution weight calculated in the part without objects. The value in this paper is 0.5.

Central coordinate loss $loss_{xy}$ is defined as equation (2):

$$loss_{xy} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} [(x_i - \hat{x}_i^j)^2 + (y_i - \hat{y}_i^j)^2] \quad (2)$$

Formula (2) represents that when the j-th detection frame of the i-th grid is responsible for a real target, the boundary frame generated by the detection frame is compared with that of the real target, and the loss of central coordinates is calculated. Where I_{ij}^{obj} indicates whether the j-th detection frame of the i-th grid is responsible for the target. If so, I_{ij}^{obj} = 1, otherwise I_{ij}^{obj} =0.

Wide and high coordinate loss is defined as:

$$loss_{wh} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} [(\sqrt{w_i^j} - \sqrt{\hat{w}_i^j})^2 + (\sqrt{h_i^j} - \sqrt{\hat{h}_i^j})^2]$$

(3)

Formula (3) represents that when the j-th detection frame of the i-th grid is responsible for a real target, the boundary frame generated by the detection frame is compared with the boundary frame of the real target, and the width and height coordinate loss is calculated.

Confidence loss $loss_{conf}$ is defined as equation (4):

$$loss_{conf} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} [\hat{C}_i^j \ln C_i^j + (1 - \hat{C}_i^j) \ln(1 - C_i^j)] +$$

$$\lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{noobj} [\hat{C}_i^j \ln C_i^j + (1 - \hat{C}_i^j) \ln(1 - C_i^j)]$$

(4)

The first term in formula (4) represents the confidence error of the object detection frame. The second term indicates that there is no confidence error of object detection frame. Where I_{ij}^{noobj} indicates that the j-th detection box of the i-th grid is not responsible for the target. $\hat{C}_i^j$ is confidence. During training, $\hat{C}_i^j$ represents the true value. The value of $\hat{C}_i^j$ is determined by whether the detection box of the cell is responsible for predicting a certain target. If so, $\hat{C}_i^j$ =1, otherwise, $\hat{C}_i^j$ =0.

$loss_{cls}$ are defined as formula (5):

$$loss_{cls} = \sum_{i=0}^{S^2} I_{ij}^{obj} \sum_{c \in classes} [\hat{P} \ln P_i^j + (1 - \hat{P}) \ln(1 - P_i^j)]$$

(5)

Formula (5) indicates that only when the j-th detection frame of the i-th grid is responsible for a real target, the boundary frame generated by the detection frame will calculate the classification loss function. The proposed network structure is shown in figure 1. Figure 2 is the spatial pyramid pooling (SPP) module [23,24], and figure 3 is the flow chart of the algorithm in this paper.

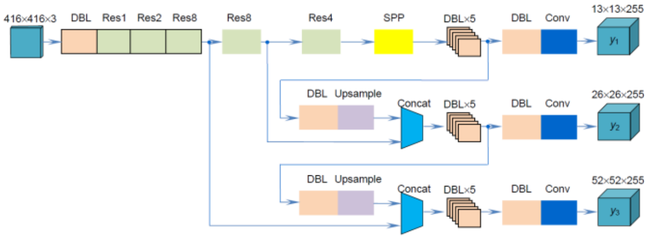


Figure 1. Modified YOLOv3 network structure

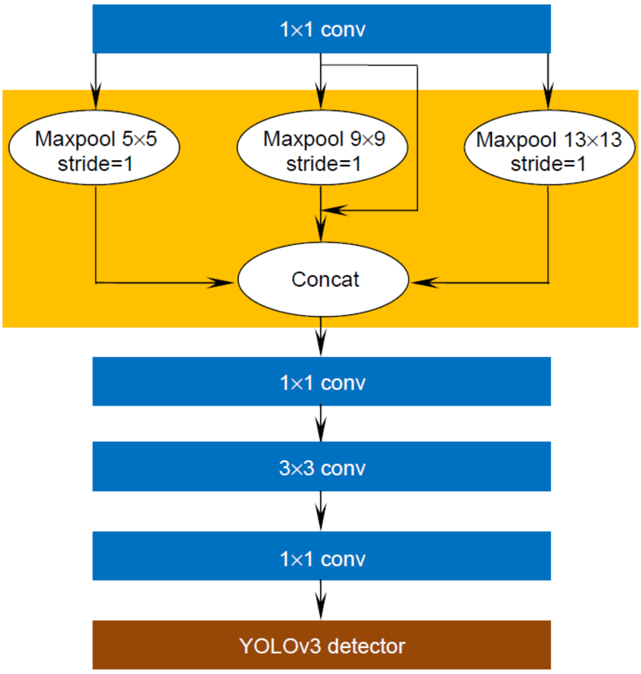


Figure 2. SPP module

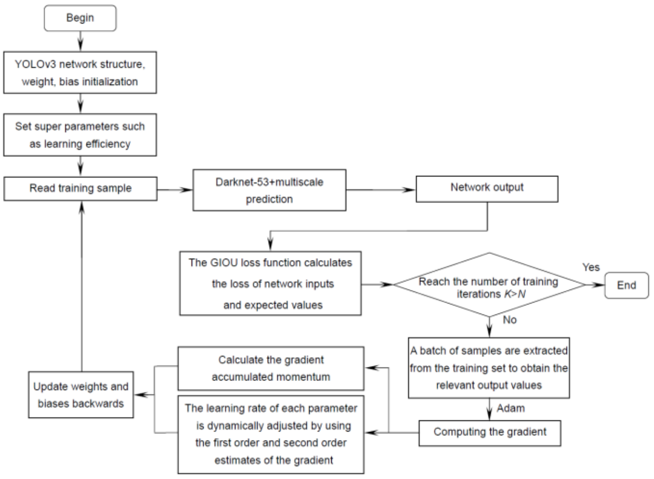

Begin
YOLOv3 network structure, weight, bias initialization
Set super parameters such as learning efficiency
Read training sample
Darknet-53+multiscale prediction
Network output
The GIOU loss function calculates the loss of network inputs and expected values
Reach the number of training iterations K>N
Yes
End
No
A batch of samples are extracted from the training set to obtain the relevant output values
Adam
Computing the gradient
Calculate the gradient accumulated momentum
The learning rate of each parameter is dynamically adjusted by using the first order and second order estimates of the gradient
Update weights and biases backwards

Figure 3. Algorithm flow chart

In YOLOv3, the mean square error (MSE) is used as the loss function to regression the center point, width and height of BBox, but in this way, the coordinate value of each point of BBox is regarded as an independent variable, without considering the integrity of the object frame, and the l_n -norm is sensitive to the scale of the object. In order to solve the above problems, Yu et al. [25] proposed to replace the traditional MSE loss function with IOU loss function according to the intersection ratio between IOU and BBox. Later, when the real frame and the prediction frame do not coincide with each other in IOU loss function, the loss is 0 and gradient return cannot be implemented. In reference [26], it is intuitively shown that the qualities of the detection results are different for the same l-norms value. Therefore, GIOU and GIOU loss functions are proposed, which are expressed as follows:

$$R_{GIOU} = O_{IOU} - \frac{A^c - U}{A^c} \tag{6}$$

$$L_{GIOU} = 1 - R_{GIOU} \tag{7}$$

Where A^c represents the area of the minimum box containing BBox and GT. U represents the total area of BBox and GT. GIOU, like IOU, can also be regarded as a distance measure, which meets the basic requirements as a loss function and has scale invariability. Meanwhile, it improves the situation that the gradient is 0 caused by the disintersection between the prediction frame and the real frame that exists in IOU as a loss function.

$$\lim_{\frac{|A\cap B|}{|C|}\to 0} R_{GIOU}(A,B) = -1 \tag{8}$$

According to equation (8), this method can effectively improve the inaccurate positioning of YOLOv3.

In addition, in order to further improve YOLOv3 to the characteristics of the power of expression, referenced by the ideas of the space pyramid in YOLOv3 SPP module is added to the network, make the original for the characteristics of the target detection figure after SPP, global features and local features to fusion, enrich the expression ability of feature maps, expanded the figure characteristics of the receptive field, to detect in the image. Standard size span is relatively large. If directly input YOLOv3 for training, it is easy to lead to over-fitting. However, after the difference between RGB and infrared images is reduced through the pre-processing of data set, the weight after vehicle target detection on RGB images is used as the initial weight of vehicle target detection, which can not only reduce the network's requirements on data volume, but also reduce the time of network training.

3. Vehicle attribute recognition network

The deeper network denotes the higher accuracy of target detection identification, but it has the longer corresponding detection time. In the ordinary static image recognition, the influence of the time factor is not very prominent, but in the video application, it needs to consider the real-time of video conditions. In the process of a frame video target detection identification, time factor is an important standard. In this experiment, including the two attribute categories of vehicle type and vehicle color, considering the different color distribution areas of different types of vehicle, such as the window and the hood of the front, the whole image area of the vehicle are as the ROI of the vehicle information for color training, which will conflict with the ROI of vehicle type. In this experiment, the vehicle type and vehicle color attributes are graded training, which can not only use different network structures, but also avoid the problem of model and vehicle color ROI conflict. Using a different network structure between the two modules during the training time, and then integrating calls to the model, which can greatly improve the accuracy and detection time of vehicle attribute recognition.

Figure 4 is the vehicle attribute recognition algorithm of this experiment. In figure 4, the input image is first re-sized to 416×416. Then CNN is run on the image for feature extraction. Finally, the threshold value of the detection result is set through the confidence degree of the model to screen the detection frame.

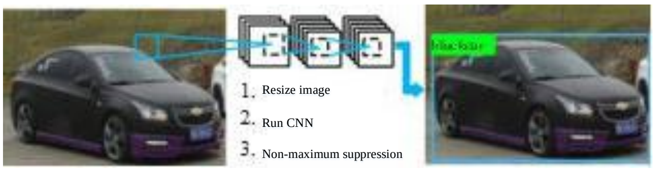


Figure 4. Vehicle attribute recognition algorithm

In this experiment, YOLOv3 neural network is used to establish the vehicle attribute recognition model. Figure 5 is the structure diagram of the vehicle type recognition network.

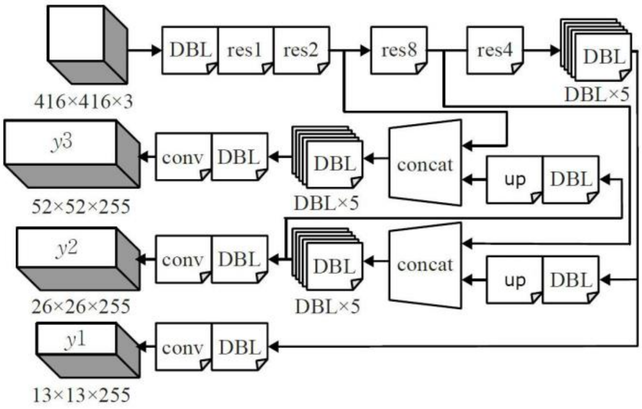


Figure 5. Vehicle type recognition network structure diagram

The details of DBL, resn and concat in figure 5 are as follows.

DBL: is the basic component of this network architecture, namely, combination of convolution +Batch Normalization +Leaky Relu.

resn: n is number, res1, res2,.., res8, indicating that an res_block contains n res_units.

concat: tensor splicing, splicing the upper samples of the middle layer and a later layer, so that the network can learn deep and shallow features at the same time, so that the expression effect is better.

up: up-sampling.

The network extracts features through several DBL components and res residual units, and then learns features of each layer through tensor splicing and fusion of up-sampling of the middle layer and a subsequent layer through concat. Finally, outputs of three scales are created, namely [y1, y2, y3] in figure 5. The bottom layer information contains global features and the middle layer information contains local features. Such splicing can give consideration to both. In addition, the idea of Feature Pyramid Networks (FPN) [27,28] is used to detect objects of different sizes with multiple scales. The finer grid element denotes the finer object that can be detected.

In the vehicle type recognition module, because need acquisition of feature points is more, the deeper the characteristic sampling network, the more models feature points collected, classification accuracy is higher, the vehicle color recognition module, vehicle color distribution is evener, its feature points is less, only need to calculate the color pixels in each ROI, characteristics of sampling network just need some simple layer can be realized. Based on the YOLOv3 neural network, some of its convolutional layers are retained, and a pooling layer is added after each convolutional layer, and then a new network structure of 23 layers is reassembled by tensor splicing. The vehicle color samples are trained to improve the overall time of vehicle attribute recognition. Figure 6 shows the improved network structure of vehicle color recognition. Where, DP is the combination of convolution layer and pooling layer. In this network, feature extraction is carried out through multiple DP components and several convolution layers. Then, tensor splicing and fusion are also carried out through concat to learn attribute features. Finally, outputs of two scales are created, namely [y1,y2] in figure 6.

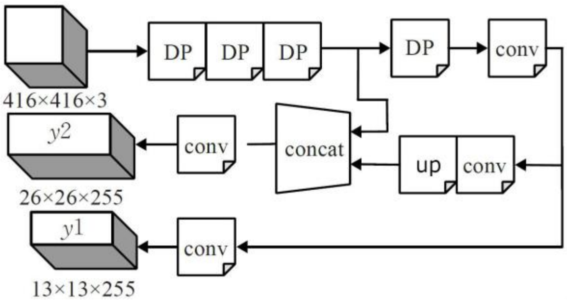


Figure 6. Vehicle color recognition network structure diagram

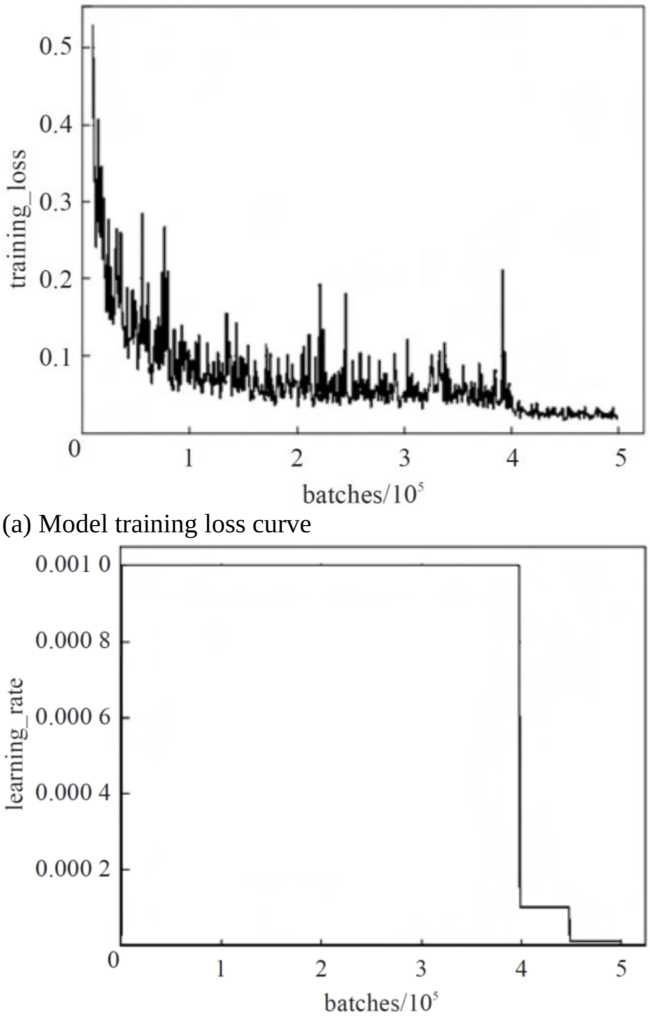
The DP combination method simplifies the network structure, reduces the network depth, and greatly shortens the time of model detection and recognition without affecting the accuracy of recognition. Combined with the model recognition model, the accuracy of vehicle attribute recognition can be improved and the overall time of vehicle recognition can be shortened by calling the vehicle color model for color recognition when the model is detected by the network.

4. Experiments

4.1. Model training

Before model training, it is necessary to prepare the vehicle image sample data set required by model training for feature learning by network. In addition to the original image sample data set, the sample region of interest required by network training, namely the label data set of the sample, should also be prepared. In this label data set, the training data need to be manually marked with the favorable region of interest and category name, which can not only improve the accuracy of the region of interest, but also reduce the interference of noise and improve the effectiveness of feature extraction. In this paper, vehicles are divided into bus, car, coach, truck, suv and van. According to the color, it is divided into black, blue, gray, green, orange, purple, red, white, champagne, yellow and silver gray.

The experiment is trained on GTX 1080Ti GPU. Before the training, some parameters needed to be adjusted. The training data set was divided into 64 batches, and the number of each batch was set to 4, so as to reduce the GPU burden and iterate at the fastest speed. The higher the learning rate parameter value is set in the training process, the higher the recognition accuracy of the obtained model will be [29-31]. However, this parameter cannot be set at will. Too high learning-rate will lead to learning bias of the network only learning data sets, so learning rate is used in this experiment is set to 0.001. In the training process, the curve changes of the model's training loss and model learning rate are shown in figure 7. Figure 7 (a) is in the process of model training curve of average loss, the average loss in the process of training the more down to studying the better the results of the model, figure 7 (b) is in the process of model training vector graph, the curve reflects the training ability of the model in the process of learning to the attribute of the size, expectations towards training before the set value, namely 0.001. As can be seen from figure 7, the training loss value of the model finally decreases to 0.02, which can meet the training requirements. However, the learning rate drops sharply after 400000 iterations, indicating that the training model has an over-fitting phenomenon at this time. Therefore, combined with the average loss curve, the optimal number of iterations should end at 400. At this time, the average loss is still at the lowest, and the learning rate reaches the highest. Stopping the training at the right time can reduce the influence of over-fitting in the training process and improve the identification accuracy of the model.



(a) Model training loss curve

(b) Model learning rate curve

Figure 7. Training loss and learning rate outcomes

4.2. Data sets

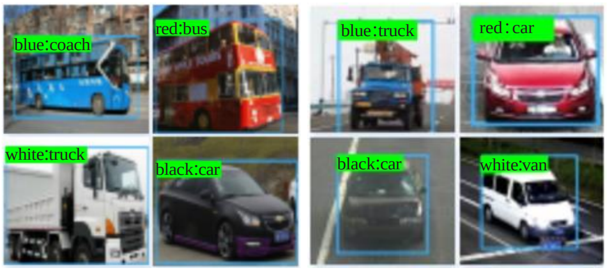
In order to realize the identification of vehicle type and color attributes, it is necessary to have sufficient training data and rich categories in the data set. In order to meet the requirements of the experiment, vehicles of different environments, different angles, different types and different colors should be included. However, in the current open vehicle data set, vehicle types are generally older, and the color is quite different from the current vehicle color, which cannot meet the data requirements of this experiment. Therefore, in order to meet the data requirements of this experiment, an AttributesCars data set containing a variety of vehicle attributes was built, including vehicles of various types and colors, totaling 20000 vehicle images, which can complete the training preparation of vehicle types and color attributes. Among them, 50% are from Stanford Cars, a vehicle data set publicly available on the Internet [32]. The types of vehicles in this data set are not much different from current vehicles, but the types of vehicles are relatively fixed, with car and suv in the majority. Therefore, the other 50% are collected from data of various types of vehicles publicly available on the Internet and data of various types of vehicles in various scenarios manually collected. In order to expand the data set and enhance the robustness of the training model, these data are zoomed and blurred respectively, among which 80% are screened one by one and labeled for training, and the remaining 20% are used for testing.

4.3. Experimental results and analysis

In the process of vehicle attribute recognition, it is necessary to meet the requirements of real-time video to the maximum extent on the premise of satisfying the recognition accuracy. Simplified in this experiment, color identification module uses a modified version of the network structure, training the model significantly reduced testing to identify the time used, and the accuracy is not affected, although in the vehicle recognition module identification for a long time, but in combination with color module module can make up for model identification of defects for a longer time, the overall vehicle identification time can satisfy the real time video requirements.

In vehicle driving, road scenes vary, and vehicle images in surveillance videos vary with the changes of road scenes. In a surveillance picture, the distance between the location of vehicles and the location of cameras directly affects whether vehicle attributes can be correctly and effectively recognized. In order to evaluate the applicability of the vehicle attribute recognition model, the test and verification are carried out in two different scenarios, close-range monitoring and traffic monitoring. In the close-range monitoring scene, the camera is close to the target position of the vehicle, and the images collected are relatively clear. The accuracy of vehicle detection and identification in the monitoring picture is relatively high. In the traffic monitoring scene, the camera is generally located in a higher position and far from the vehicle, so the pixel deviation of the vehicle image collected is larger than that of the close shot, the target is smaller, and the accuracy is relatively low.

Figure 8 is the recognition effect diagram of the training model in this experiment. The top of the target vehicle in the figure shows the recognition results of the color and type attributes of the vehicle. Figure 8(a) is the recognition result of some vehicle attributes in the close-range monitoring scene. The vehicle image in the figure is large with clear color and the recognition result is correct. Figure 8(b) is the attribute recognition result of some vehicles in the traffic monitoring scene. The vehicle image in the figure is small, the color is barely visible, and the recognition result is correct. As can be seen from figure 8, vehicle types and colors can be correctly identified in both close-range monitoring scenarios and traffic monitoring scenarios, proving the feasibility of this model.



(a) Identification results of close-up surveillance (b) Traffic monitoring identification results

Figure 8. Model recognition results

In experimental results can be seen that the experimental model can correctly identify the type of vehicle under different scenarios and color attributes, for verifying the superiority of the method in the vehicle properties recognition, the method of this article and the other in models and research methods of vehicle color recognition, because this study involves two vehicle types and color properties, and is a classification to the training. The vehicle type and color attributes are compared respectively. A comparative analysis is made on the recognition accuracy and recognition time respectively. The comparative analysis of the recognition results of vehicle type and vehicle color is shown in Table 1 and Table 2 respectively.

Table 1. Vehicle type identification results

Method	Accuracy/%	Time/ms
Capsule network	91.38	44.38
Improved CNN	94.99	26.29
YOLOv3	97.73	38.67

Proposed	98.18	31.99

Table 2. Vehicle color recognition results

Method	Accuracy/%	Time/ms
SVM	87.63	25.87
Lighting treatment	89.76	31.69
YOLOv3	93.47	38.69
Proposed	93.19	3.85

As can be seen from Table 1, on the premise of 91.38% accuracy of the capsule network-based method, the vehicle identification time is 44.38ms. The accuracy of the improved CNN method is 94.99%, and the recognition time is reduced to 26.29 ms. The accuracy of the original YOLOv3 method is 97.73%, and the corresponding recognition time is increased to 38.67ms. Moreover, it is not comprehensive enough to recognize a few vehicle categories.

The accuracy of the proposed method is 98.18%, and the recognition time is 31.99ms. Compared with the original YOLOv3 method, the recognition time is slightly reduced under the premise of improved accuracy.

As can be seen from Table 2, the accuracy rate of vehicle color recognition based on SVM method is 87.63%. The accuracy of vehicle color recognition based on illumination processing method is 89.76%. However, YOLOv3 method can achieve 93.47% accuracy and 38.69ms recognition time, which is significantly shorter than other methods. The accuracy of the method proposed in this paper is 93.19%, and the recognition time is only 3.85ms. Compared with the YOLOv3 method, the recognition time is greatly shortened without affecting the recognition accuracy, and it is more practical for the application of vehicle recognition in video.

Based on the results of Table 1 and Table 2, it can be seen that the vehicle multi-attribute recognition model proposed in this paper can effectively recognize the vehicle color while recognizing the vehicle type, and the model recognition time is 31.98 ms.

The recognition time combined with vehicle color is 3.85ms. The overall vehicle recognition time is 35.84ms, which can meet the requirements of real-time video. In addition, the combination of vehicle type and color category makes the vehicle recognition results more comprehensive, with an average accuracy of 95.63%, which can be more effectively applied to the detection and recognition of vehicle attributes in real traffic.

Figure 9 shows the detection results of some vehicle data sets. It can be seen from the comparison between column 9(a) and column 9(b) that when the original YOLOv3 network detects infrared targets, its detection ability is poor for large targets at short distance or edge targets, and there are problems of false detection and missed detection when the target is small. The author of YOLOv3 also pointed out that YOLOv3 network could detect medium and large targets [33,34]. The comparison of column 9(a) with column 9(c) and column 9(d) shows that the detection ability of the network after modification has been significantly improved for large targets in short distance and edge targets, and the accuracy of positioning has also been improved to some extent. The box of network prediction is closer to the real value. By comparing the columns in figure 9(c) and figure 9(d), it can be seen that the network error detection and error detection capability are reduced after the addition of SPP module. After overcoming the original deficiency of YOLOv3, the target detection accuracy can be further improved compared with the method of only modifying GIOU loss.

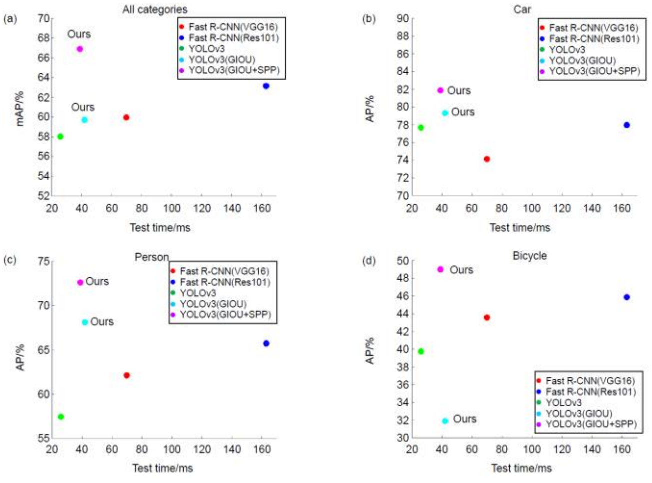


Figure 9. (a) The results of detection speed and accuracy of all categories of different networks; (b) The results of vehicle detection speed and accuracy of different networks; (c) The results of human detection speed and accuracy of different networks; (d) The results of bicycle detection speed and accuracy on different networks

5. Conclusion

This paper proposes a vehicle multi-attribute recognition method based on deep neural network in multi-scene. First on real road scenario monitoring image, collection contains various types and various colors of vehicle data sets, and classify the screening and label processing, training after the improved YOLOv3 neural network to get the vehicle multiple attribute recognition model, this model can be achieved on the premise of shortening the recognition time and high recognition accuracy, and the testing requirement with good effect, applicable to all monitoring images in real road scenarios. In this paper, the sample data sets of vehicle type and color attributes are graded to effectively improve the accuracy and real-time performance of attribute classification. Vehicles are obtained by test data set for training multiple attribute recognition models is tested, and the experimental results show that the proposed approach in different road scenario is of high precision, meet the requirement of practical use, and can effectively identify the types of vehicles and color information, and has high accuracy and good applicability and robustness.

Acknowledgements.

This paper is supported by Science and Technology Development Plan of Henan Province in 2019. Yuke 2019 Project No. 16. Fund No. 192102210262.

References

[1] Damaj I, Khatib S A, Naous T, et al. Intelligent Transportation Systems: A Survey on Modern Hardware Devices for the Era of Machine Learning[J]. Journal of King Saud University - Computer and Information Sciences, 2021.

[2] A. Chowdhury, G. Karmakar, J. Kamruzzaman and S. Islam, "Trustworthiness of Self-Driving Vehicles for Intelligent Transportation Systems in Industry Applications," in IEEE Transactions on Industrial Informatics, vol. 17, no. 2, pp. 961-970, Feb. 2021, doi: 10.1109/TII.2020.2987431.

[3] Li H, Yin S, Liu J, et al. Novel gaussian approximate filter method for stochastic non-linear system[J]. International Journal of Innovative Computing, Information and Control. 13(1): 201-218, 2017.

[4] Shoulin Yin, Ye Zhang, Shahid Karim. Large Scale Remote Sensing Image Segmentation Based on Fuzzy Region Competition and Gaussian Mixture Model[J]. IEEE Access. volume 6, pp: 26069 - 26080, 2018.

[5] Lai, Pan-Huang, Guo, et al. Location-aware fine-grained vehicle type recognition using multi-task deep networks[J]. Neurocomputing, 2017.

[6] Chen Q, Liu W, Yu X. A Viewpoint Aware Multi-Task Learning Framework for Fine-Grained Vehicle Recognition[J]. IEEE Access, 2020, 8:171912-171923.

[7] Zhao W, Ye J, Yang M, et al. Investigating Capsule Networks with Dynamic Routing for Text Classification[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

[8] Chen C, Cai X, Zhao Q, et al. Vehicle type recognition based on multi-branch and multi-layer features[C]// 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2017.

[9] Yin Shoulin, Liu Jie, Li Hang. A Self-Supervised Learning Method for Shadow Detection in Remote Sensing Imagery[J]. 3D Research, vol. 9, no. 4, December 1, 2018. EI(JA) https://doi.org/10.1007/s13319-018-0204-9

[10] Teng Lin, Hang Li and Shoulin Yin. Modified Pyramid Dual Tree Direction Filter-based Image De-noising via Curvature Scale and Non-local mean multi-Grade remnant multi-Grade Remnant Filter [J]. International Journal of Communication Systems. v 31, n 16, November 10,pp. e.3486.1-e.3486.12, 2018.

[11] Shahid Karim, Ye Zhang, Shoulin Yin, Muhammad Rizwan Asif. An Efficient Region Proposal Method for Optical Remote Sensing Imagery[C]. IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, pp: 2455-2458, July 2018.

[12] Choi S H, Yun J P, Sim S B, et al. Edge-based text localization and character segmentation algorithms for automatic slab information recognition[C]// International Conference on Image Analysis & Signal Processing. IEEE, 2010.

[13] Yin Shoulin, Liu Jie, Teng Lin. A new krill herd algorithm based on SVM method for road feature extraction[J]. Journal of Information Hiding and Multimedia Signal Processing, v 9, n 4, p 997-1005, July 2018.

[14] X. Xue, J. Ding and Y. Shi, "Research and application of illumination processing method in vehicle color recognition," 2017 3rd IEEE International Conference on Computer and Communications (ICCC), 2017, pp. 1662-1666, doi: 10.1109/CompComm.2017.8322822.

[15] Hang R, H Sun. Vehicle Multi-attribute Recognition Based on Faster R-CNN[J]. Computer Technology and Development, 28(10), 129-134, 2018.

[16] Shoulin Yin, Hang Li, Asif Ali Laghari, et al. A Bagging Strategy-Based Kernel Extreme Learning Machine for Complex Network Intrusion Detection[J]. EAI Endorsed Transactions on Scalable Information Systems. 21(33), e8, 2021. http://dx.doi.org/10.4108/eai.6-10-2021.171247

[17] Qingwu Shi, Shoulin Yin, Kun Wang, et al. Multichannel convolutional neural network-based fuzzy active contour model for medical image segmentation. Evolving Systems (2021). https://doi.org/10.1007/s12530-021-09392-3

[18] Jisi A and Shoulin Yin. A New Feature Fusion Network for Student Behavior Recognition in Education [J]. Journal of Applied Science and Engineering. vol. 24, no. 2, pp.133-140, 2021.

[19] Technicolor T, Related S, Technicolor T, et al. ImageNet Classification with Deep Convolutional Neural Networks networks[C]// Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012.

[20] Ding J, Yu M, Zhu L, et al. Diverse spectral band-based deep residual network for tongue squamous cell carcinoma

classification using fiber optic Raman spectroscopy[J]. Photodiagnosis and Photodynamic Therapy, 2020, 32:102048.

[21] Yang Z, Papanikolaou S, Reid A, et al. Learning to Predict Crystal Plasticity at the Nanoscale: Deep Residual Networks and Size Effects in Uniaxial Compression Discrete Dislocation Simulations[J]. Scientific Reports, 2020, 10(1):8262.

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

[23] Z. Ke, C. Le and Y. Yao, "A multivariate grey incidence model for different scale data based on spatial pyramid pooling," in Journal of Systems Engineering and Electronics, vol. 31, no. 4, pp. 770-779, Aug. 2020, doi: 10.23919/JSEE.2020.000052.

[24] S. Yin and H. Li. Hot Region Selection Based on Selective Search and Modified Fuzzy C-Means in Remote Sensing Images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 5862-5871, 2020, doi: 10.1109/JSTARS.2020.3025582.

[25] Yu J, Jiang Y, Wang Z, et al. UnitBox: An Advanced Object Detection Network[C]\\ Proceedings of the 24th ACM international conference on MultimediaOctober 2016 Pages 516–520. https://doi.org/10.1145/2964284.2967274

[26] Rezatofighi H, Tsoi N, Gwak J Y, et al. Generalized intersection over union: a metric and a loss for bounding box regression[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[27] Y. Li, X. Pei, Q. Huang, L. Jiao, R. Shang and N. Marturi, "Anchor-Free Single Stage Detector in Remote Sensing Images Based on Multiscale Dense Path Aggregation Feature Pyramid Network," in IEEE Access, vol. 8, pp. 63121-63133, 2020, doi: 10.1109/ACCESS.2020.2984310.

[28] Shoulin Yin, Hang Li, Desheng Liu and Shahid Karim. Active Contour Modal Based on Density-oriented BIRCH Clustering Method for Medical Image Segmentation [J]. Multimedia Tools and Applications. Vol. 79, pp. 31049-31068, 2020.

[29] Xiaowei Wang, Shoulin Yin, Hang Li. A Network Intrusion Detection Method Based on Deep Multi-scale Convolutional Neural Network[J]. International Journal of Wireless Information Networks. 27(4), 503-517, 2020.

[30] Shoulin Yin, Jie Liu, and Lin Teng. A Sequential Cipher Algorithm Based on Feedback Discrete Hopfield Neural Network and Logistic Chaotic Sequence [J]. International Journal of Network Security. Vol. 22, No. 5, pp. 869-873, 2020.

[31] Yin, S., Li, H. & Teng, L. Airport Detection Based on Improved Faster RCNN in Large Scale Remote Sensing Images [J]. Sensing and Imaging, vol. 21, 2020. https://doi.org/10.1007/s11220-020-00314-2

[32] Lin Wang, Jiangyun Cai. Research on Deep Neural Network in Multi-scene Vehicle Attribute Recognition [J]. Computer Engineering and Applications. 57(9), 2021. (In Chinese)

[33] Xiaowei Wang, Shoulin Yin, Desheng Liu, et al. Accurate playground localisation based on multi-feature extraction and cascade classifier in optical remote sensing images [J]. International Journal of Image and Data Fusion, vol. 11, no. 3. pp. 233-250, 2020.

[34] Xiaowei Wang, Shoulin Yin, Ke Sun, et al. GKFC-CNN: Modified Gaussian Kernel Fuzzy C-means and Convolutional Neural Network for Apple Segmentation and Recognition [J]. Journal of Applied Science and Engineering, vol. 23, no. 3, pp. 555-561, 2020.