IOP Conference Series: Materials Science and Engineering

PAPER • **OPEN ACCESS**

Research on visual vehicle detection and tracking based on deep learning

To cite this article: Yaoming Zhang *et al* 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **892** 012051

IOP Conf. Series: Materials Science and Engineering **892** (2020) 012051 doi:10.1088/1757-899X/892/1/012051

Research on visual vehicle detection and tracking based on deep learning

Yaoming Zhang[1,3], Xiaoli Song[1], Mengen Wang[1], Tian Guan[1], Jiawei Liu[1], Zhaojian Wang[1], Yajing Zhen[1], Dongsheng Zhang[2] and Xiaoyi Gu[1]

[1] School of Automobile, Chang ' an university,710064,Xi'an Shaanxim,China;
[2] Testing department, Ningbo Traffic Construction Engineering Testing Center Co.,Ltd, Ningbo,Zhejiang Province,China

[3] Email: 924565747@qq.com

Abstract. At present, vehicle tracking technology is still a hot research direction all over the world. The main work of this paper is divided into two parts: we should complete visual vehicle detection based on deep learning and vehicle tracking based on similarity measurement and association algorithm. The main work of this paper is as follows: Based on the YOLOV3 detection algorithm, this paper puts forward some new ideas and ideas, which are applied to the video detection and experimental results. Based on the structure of YOLOV3 convolutional neural network, the vehicle video detection experiment is carried out under the framework of TensorFlow, and the detection results are evaluated by the detection accuracy and error detection rate. Firstly, the augmented data set of video detection is made, and a section of driving record video is recorded under the surrounding road conditions of the school. After that, the data set is used to train the weightless network of YOLOV3. Finally, the test results between KITTI data set and augmented data set are compared. Then the network is used to compare the detection results of different detection algorithms. The second work is to build the vehicle tracking model after getting the object selection frame after video detection. In this model, the classic hog feature extraction method is used to extract the vehicle appearance features, and then the motion similarity is calculated. After getting the total similarity between the targets in the front and back images, the target trajectory is recovered or released by the target trajectory management algorithm to achieve the tracking effect. The above vehicle tracking model is optimized by three evaluation parameters: MOTA, MOTP and IDS. In the end of this paper, the tracking effect experiments under different detection algorithms, different detection thresholds and the peeling analysis of tracking features are carried out for the vehicle tracking part, and some conclusions are obtained.

1. Preface

At present, in most of the image processing and detection algorithms, it is difficult to achieve the real-time requirements of real vehicle video detection. In the vehicle detection algorithm, the end-to-end detection algorithm based on deep learning can recognize the vehicle in the video quickly and Simultaneously [1]. Compared with the feature classification algorithm, the resolution and processing method of the image are changed. The use of the algorithm also shortens the time cycle of deep learning. Vehicle detection is the basis of vehicle tracking and license plate recognition. A good vehicle detection algorithm can provide a powerful guarantee for the latter.

Vehicle video detection and tracking is one of the research contents in the field of computer vision. With the rapid development of artificial intelligence, vision-based vehicle detection and tracking has become an irreplaceable part of driverless vehicle technology [2]. Advanced and accurate vehicle detection technology and vehicle tracking are bound to promote the further development of driverless vehicles. The main function of vehicle detection technology can accurately locate and detect vehicles in driving video or pictures. On the basis of vehicle detection, vehicle tracking distributes the ID information of the existing detection results, so that each detected vehicle target can be connected with a coherent trajectory in the video.

At present, there are two widely used methods in the field of target detection based on deep learning. One is the method of using deep learning network to classify the feature areas after the candidate areas are generated based on images, mainly R-CNN series [3]. The other is based on the end-to-end processing of the image detection method, mainly YOLO series, the latter is slightly less accurate than the former, but the detection speed has reached the requirements of real-time detection [4].

2. Video Vehicle detection based on deep learning

Multi-target recognition algorithm based on deep learning can be generally regarded as multi classification problem, and the main task is to separate the target and image background. Before the deep learning theory is widely used, the general multi-target detection is carried out in three steps, the first is the feature extraction of the object, the second is the classification of the object, and the last is the combination of the two in the detection process [5].

There are many frameworks for in-depth learning, and the TensorFlow framework is selected in this paper. The TensorFlow framework accelerates the construction and architecture of deep learning neural network [6]. The TensorFlow framework provides a series of network architectures to implement many basic functions, such as speech recognition and limb motion recognition. The computational flow in the TensorFlow framework requires a computational static graph to be constructed for computations. The YOLOV3 video vehicle detection algorithm used in this paper is based on the TensorFlow framework [7].

The YOLO series of algorithms are representative algorithms in end-to-end deep learning. They deal with object detection as a problem of regression analysis. Compared with the R-CNN series algorithm, the R-CNN series algorithm generates thousands of image candidate boxes, then uses the classifier and detector to determine if each candidate box contains the object to be detected and the category of the object to be detected. The entire YOLOV3 network structure does not use pooling layer and full connection layer, but uses the structure of residual blocks, which greatly reduces the training time of deep learning network. In addition, the change of tensor size in YOLOV3 is achieved by changing the convolution kernel step [8]. In YOLOV2, the maximum pooled layer is used to change the size of the tensor, which is reduced by half five times [9].

The network structure of the current algorithm network consists of the following parts: convolution layer, residual layer, and full connection layer.

3. Vehicle video tracking based on similarity measurement and association algorithm

3.1. Appearance feature extraction based on hog feature

Hog feature is a kind of multiple information feature set, which can not only describe the contour of vehicle, but also extract vehicle motion information. The hog feature was applied in the static image of pedestrian detection, and then gradually used in the vehicle detection of video image. The problem of driverless vehicle tracking can be solved by extracting the appearance features of the vehicle using the hog features [10].

The edge of the object in the image can be represented by the density and gradient of the gray value of the pixel in the edge distribution. Firstly, the image is transformed into gray image, and then the image is segmented. The smallest cell is cell. There are a certain number of pixels in the cell. In the

general process of hog feature extraction, the gradient histogram of the whole block is constructed by constructing the gray level gradient histogram of the cell pixel points. The features of each block are fused to generate the hog feature descriptor. Figure 1 shows the process of hog extraction. At the same time, normalization also reduces the interference of shadow and light changes in the image.

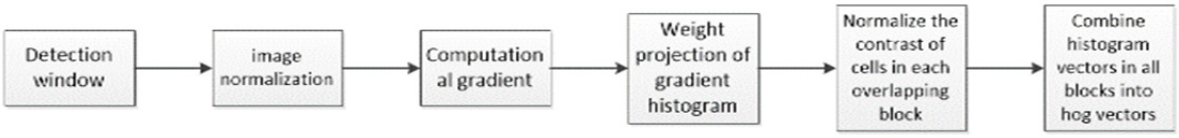

Figure 1. Hog feature extraction process.

When compared the hog features of the two images before and after the hog feature vector, we can get the appearance similarity. Because the hog features are all vector patterns, there are many ways to define the appearance similarity based on the hog feature. This paper selects another method to calculate the angle between the two hog feature vectors to define the appearance similarity [11]. At the same time, the cosine function is used to normalize the angle to obtain the cosine value. The appearance similarity is calculated as follows. Define μ_1 as the appearance characteristic value. Define x_i as the hog character of the i frame. Define θ as the angle between the front and back frames.

$$\mu_1 = cos(x_i, x_{i+1}) = cos(\theta) \quad (1)$$

3.2. Geometric feature extraction based on IOU

IOU is the overlapping rate of the generated candidate frame and the original marker frame, which is calculated as shown in formula 2. The ratio is 1 when the candidate frame and the original marker frame overlap completely.

$$IOU(c, g) = \frac{area(c) \cap area(g)}{area(c) \cup area(g)} \quad (2)$$

The detection idea here is very similar to Canny algorithm's double threshold detection. The high threshold target will cover the target of the low threshold target in the next frame. Such target matching does not need to detect all the frames on the video time series axis at one time, but only needs to read the current frame and the next frame, which more truly reflects the real tracking process. However, if the deformation of the object is large in the tracking process, the tracking effect will be affected.

Calculate the geometric feature similarity of detected and tracked vehicles. Select the prediction box of two consecutive pictures, then we can obtain the cross area between the prediction boxes of two pictures. Finally, the geometric characteristics of the video vehicle are normalized, and the geometric similarity is calculated by dividing the cross area by the total area.

Calculate the size similarity between vehicle and detection vehicle. The definition of size similarity is as follows, which is the one-dimensional feature of the two image prediction frames. After the geometric information of the prediction frame of the next frame is obtained by Kalman filter equation, the width and height of the prediction frame of the first and second frames can be calculated.

3.3. Extraction of motion feature information based on Kalman filtering principle

Kalman filter theory is based on probability theory. In this theory, Kalman state equation is put forward, through which the system model is established. The state equation of input and output can be used to estimate and predict the state of the target at the next moment. Through this state, the movement and time state of the target are updated continuously [12].

Kalman filter updates and feeds back the state of the target mainly through two steps. Firstly, it forecasts the target state, and then obtains the next prediction and estimation through the analysis of the existing current state and the error [13]. After that, the Kalman filter equation is modified to analyze the existing estimation and observation value in the next period to get the accurate estimation, update the previous prediction state, and repeat the call, which is the Kalman filter workflow.

4. Simulation experiment and result analysis

4.1. Comparison of experimental results based on YOLOV3 model

In order to know the difference of training effect between the data set made in this paper and the KITTI data set on the data set of YOLOV3, this paper made the data set by ourselves, applied the initial weight network of YOLOV3 to experiment, and compared the results of the network model trained by the two data sets through the average accuracy of the target detection evaluation index.

In this paper, we use the vehicle driving data collected by ourselves to mark and make the data set, and then merge it with the KITTI data set to get the expanded data set. Firstly, we install Python 3.7 and Anaconda 3.5 under windows to complete the installation of labeling in the environment of windows system, and start labeling. A total of 1400 road condition pictures are sampled in the data set, each picture is 1280 * 960 in size. 900 pictures in the data set are selected as the training set, 200 as the training verification set and 300 as the test set. The data set effectively represents the complex working conditions in the city. Up to 5 vehicles, pedestrians and bicycles can be seen in each image, which also includes various degrees of occlusion, deformation and dislocation.

We uses the accuracy of detection and the false detection rate as the model of the algorithm [14].

$$AP = \frac{TP}{TP+FP} = \frac{TP}{Toallpositive} \tag{3}$$

Where true positives (TP) is the number of correct detections in the detection, that is, the positive number. False positives (FP), is the number of false detections. Objects that were not originally targets of this class are divided into the number of targets of this class. The experimental results are compared with the target detection results of YOLOV3 trained by KITTI data set [15].

Using the augmented training verification set which contains 900 new pictures made in this paper to train to get the weight file of the YOLOV3 network in this data set and under this data set, adding 200 new pictures in the data set to the original verification set, and then 300 new pictures are added to the test set. Compared with the training of KITTI data set, the experimental results are as follows:

It can be seen from the experimental results that with the increase of data set, the detection accuracy of neural network for the target will increase. The recognition accuracy of the deep learning neural network under the training of the expanded data set is 0.82% higher than that of the deep learning neural network under the training of the KITTI data set itself, and the false detection rate is reduced by 0.26%. The larger the amount of training data, the better the detection effect of deep learning neural network, with the deepening of deep learning network and the increasing amount of training data. Generally speaking, the more training data, the longer training time.

4.2. Target detection experiments with different computational models

In this paper, different detection algorithms are compared for the target detection experiment. Because this experiment is based on the video vehicle target detection. This part of the experiment uses a different evaluation standard, using the average accuracy and video transmission rate as the indicators of the comparative evaluation target detection method [16]. In this paper, Fast R-CNN, YOLOV2, YOLOV3 three target detection algorithms are tested, all under the TensorFlow framework, and the training is carried out with the data set made in this paper.

Table 1. Comparison of Fast R-CNN, YOLOV2, YOLOV3 detection results.

Algorithm	Input	V(f·s^{-1})	AP/% Car
Faster R-CNN	600×--	0.56	94.78
YOLOV2	416×416	114.26	66.45
YOLOV3	416×416	48.60	90.21
YOLOV3	1280×960	7.60	92.19

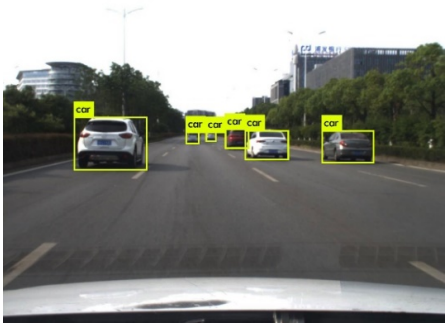

Figure 2. YOLOV3 test result.

The experimental data obtained by these algorithms are trained and tested with the data set made by ourselves. It can be seen from the Table 1 that the improved algorithm proposed in this paper obtains 92.19% detection accuracy, which is still superior to other types of methods in terms of accuracy, including the classical region segmentation detection algorithm Fast R-CNN and the end-to-end detection algorithm YOLOV3. And it achieves a speed of 7.60 in terms of transmission speed, meeting the requirements of real-time detection. And Figure 2 is the vehicle information extracted in the experiment.

4.3. Vehicle video tracking experiment under different detection algorithms

In order to determine the appropriate threshold value of target detection, we should use the YOLOV3 target detection model uniformly, set 0.4, 0.5, 0.6 three different thresholds for target detection, evaluate the experimental tracking results, apply the clear-mot standard, and analyze the best range of threshold selection. After that, the feature extraction part of the target tracking algorithm is stripped and analyzed, and the influence of three different features on the target tracking effect is analyzed. Using the YOLOV3 computing model, the tracking effect of using the appearance feature extraction based on the hog feature alone and the tracking effect of using the moving feature of the target alone are compared [17]. And the target feature and the similarity associated are evaluated. Finally, the video tracking effects of three different target detection algorithms are analyzed and compared. The target detection algorithm of YOLOV3 is applied to predict and detect the target in each frame in real time, calculate the hog appearance feature and appearance similarity of each frame, and then predict the target motion information of the next frame through Kalman filtering algorithm, calculate the motion feature, geometric feature and size feature, and get the prediction frame after Kalman filtering, which is similar to that of YOLOV3[18]. At last, we use different target detection algorithms SSD and Fast R-CNN to modify the detection part, analyze different target detection algorithms and the video tracking effect under the results, and evaluate by MOTA and MOTP.

Experimental result: In this experiment, YOLOV3 is used as the detection algorithm to detect the video made by ourselves, and the confidence setting is greater than 0.4. The results of SSD and Fast R-CNN were compared. The clear-mot evaluation results of the three algorithms in the video are shown in Table 2. The greater the value is, the better the detection result is. The smaller the value is, the worse the detection result is. The evaluation indexes include Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), and target information jump variable IDs.

Table 2. Comparison of tracking experiments based on three different multi-target detection algorithms.

Multitarget detection algorithm	MOTA↑	MOTP↑	IDS↓
YOLOV3	71.2	73.3	7
SSD	69.1	76.8	7
Faster-R-CNN	75.5	79.3	5

Due to the difference of the average accuracy and transmission rate of the target detection algorithm, different tracking algorithms will have different tracking effects. We can see that compared with YOLOV3, the MOTA of Fast R-CNN is slightly higher than that of YOLOV3. This is because the average detection accuracy of R-CNN algorithm for the target is higher than that of the Yolo series algorithm, which can also be reflected in the detection algorithm experiment, so the higher the average accuracy of the target detection, the higher the MOTA obtained from the video experiment. However, only the YOLOV3 algorithm can achieve real-time video detection, that is, the transmission rate meets the requirements of video processing, while the Fast R-CNN and SSD can not meet the requirements of video continuity. The detection of the target frame may be delayed during video tracking, resulting in inaccurate tracking position.

Table 3. Comparison of tracking effects under different thresholds.

Detection threshold	MOTA↑	MOTP↑	IDS↓
0.3	72.2	78.9	5
0.4	71.2	79.3	5
0.5	70.7	79.1	5

In Table 3, it can be seen that when the detection threshold increases more than 0.4, the number of missed detection will increase due to the increase of the requirements for detection results, and some correct targets will be missed because they are lower than the detection threshold. In the whole missed detection and false detection, the number of missed detection is the main part, so the value of MOTA will decrease with the increase of missed detection, that is, the tracking accuracy will decrease with the increase of threshold. However, it can be seen that the change of threshold does not have a great impact on MOTP. The IDS values under the three thresholds are the same, so the main detection threshold will have an impact on vehicle tracking accuracy. After trade-off, the threshold value of target detection in the tracking experiment is 0.4.

Table 4. Tracking effect after feature extraction of stripping analysis.

Target characteristics	MOTA↑	MOTP↑	IDS↓
Hog appearance features	54.5	51.2	11
IOU motion characteristics	40.1	47.3	16
Feature combination	71.2	79.3	5

In Table 4, It can be seen from the experimental results of peel off analysis that both the accuracy of vehicle tracking and the accuracy of location are greatly reduced. When only the appearance features are used as the basis for correlation, it is easy to confuse with the background or other similar scale targets. Even if the front vehicle keeps the same size in the field of vision at a constant speed, when the driving environment is light or environment color When the color changes, only the appearance features, it is difficult to distinguish the target difference. The jump of target information will be very frequent. When there are only moving features of the target, it is easy to confuse the objects with similar colors in the video. Because most of the rear images of the vehicle are very similar in size and shape, it is difficult to distinguish two vehicles with different colors without distinguishing the appearance features. The tracking effect is not ideal.

5. Conclusion

Vehicle vision detection and tracking based on deep learning is a hot research direction of computer vision. At the same time, vehicle detection and tracking is also an irreplaceable role in the development of driverless vehicles. In this paper, a vehicle detection and tracking algorithm is proposed for pedestrian detection algorithm, which can detect and track vehicles in real time, effectively improve the accuracy of feature extraction in video, and also greatly improve the detection speed. In this paper, the characteristics of vehicles are extracted and convoluted by using the YOLOV3

algorithm in the framework of tensor flow, and the predicted values are obtained by using the YOLOV3 algorithm. The prediction value is combined with the appearance feature extracted by the hog and the IOU motion feature extracted by the Kalman filter. Through the similarity association algorithm and the Hungarian algorithm, the front and rear frame targets are correlated and the matching is globally optimal. Finally, the vehicle is effectively tracked by the target management algorithm. Compared with the single detection of the vehicle tag in the video, the tracking algorithm adds ID information to each detected vehicle, which can effectively identify and continuously track the vehicle to achieve the effect of automatic cruise.

Acknowledgements
This work was supported in part by the National Natural Science Foundation of China under Grant U1664264 and in part by the Major scientific and technological innovation projects of Shandong Province under Grant 2019JZZY020904.

Reference
[1] Xiaojie Wu 2018 Research on vehicle detection and tracking based on video *Chang'an University*
[2] Yanqing Sun 2018 Research on video vehicle detection and tracking method based on depth model *ZhongbeiUniversity*
[3] Mengxia Du 2017 Research on vehicle detection based on depth learning *Xi'an University of Technology*
[4] Weiwei Zhang 2015 Video target detection and tracking for vehicle driving assistance system *Hunan University*
[5] Hao Guan 2016 Progress and prospect of deep learning in video target tracking *AC Automation Journal* **06** 841
[6] Xinmeng Jiang 2017 Application Research of convolution neural network based on TensorFlow *Central China Normal University*
[7] Fukai Zhang 2019 Fast vehicle detection method based on improved YOLOV3 *Computer E ngineering And Applications* **02** 15
[8] Dianwei Wang 2018 Improved YOLOV3 algorithm for pedestrian detection in infrared video image *Journal of Xi'an University of Posts and telecommunications* **04** 51
[9] Zhou Li Miaohua Huang 2018 Real time vehicle detection based on the YOLOV2 model China Mechanical Engineering **29** 1870
[10] Xuli Wu 2017 Research on pedestrian detection and tracking algorithm based on joint features of HOG and Haar *University of Electronic Science and technology*
[11] Jinxin Guo 2013 Face recognition based on HOG multi-feature fusion and random forest *Computer Science* **10** 280
[12] Dingcong Peng 2009 The basic principles and applications of Kalman filter *Software Guide* **11** 33
[13] Qiaona Xing 2014 Recognition and tracking of infrared landmarks based on Kalman filter *Journal of Texas A & M* **04** 28
[14] Fengbin Zhu 2018 Research and implementation of video vehicle detection algorithm based on deep learning *Hangzhou Dianzi University*
[15] Shiqi Song 2019 Vehicle Classification and tracking in complex scene based on improved YOLOV3 *Journal of the Shandong University*
[16] Qin Wan 2006 Detection and tracking of multiple moving objects in real-time video *Hunan University*
[17] Honghui Zhang 2018 Multi target tracking of complex traffic video based on deep learning *Beijing Jiaotong University*
[18] Yiyuan Shi 2004 Research on road target detection and tracking based on Kalman filter *Chang'an University*