Journal of Physics:
Conference Series

PAPER • **OPEN ACCESS**

Vehicle Detection: A Review

To cite this article: Chaochao Meng *et al* 2020 *J. Phys.: Conf. Ser.* **1634** 012107

Vehicle Detection: A Review

Chaochao Meng[1], Hong Bao[1*] and Yan Ma[1]

[1] Beijing Key Laboratory of Information Service Engineering, Beijing Union University, No.97 Beisihuan East Road, Chao Yang District, Beijing 100101, China

[*]Corresponding author's e-mail: baohong@buu.edu.cn

Abstract. Vehicle detection based computer vision is the essential algorithm in autonomous driving, aims at identifying which locating vehicles by digital images or videos. The basic idea of vehicle detection is detecting "blocks," which reflects the position of the vehicle in images or videos. Besides, this paper discusses 3D vehicle detection algorithms based on stereo perception, which originated from advanced planar vehicle detection perception. Finally, this paper summarizes the vehicle detection algorithms in recent years in terms of the difference between the feature extraction approach and the perceived results. It proposes hypotheses for further in-depth study of the vehicle detection algorithms.

1. Introduction

The World Health Organization released "the State of Global Road Safety Report 2018," on December 7, 2018. Despite the continuous improvement of road safety, up to 2018, 1.35 million people died in traffic accidents every year [1]. Road traffic accidents have caused massive trauma to families and society and also caused economic losses. The losses account for 3% of the gross national product of each country [2].

A critical factor in road traffic accidents is the traditional closed-loop control mode --- road-vehicle-person. In this control mode, the unpredictability of human behavior is the primary factor affecting the stability and security of the mode.

Society should explore a new way of the transportation system to overcome the transportation issue with technology. It is necessary to invite "people" out of the closed-loop system, and it should form the "vehicle-road." In other words, empowering cars with human perception methods instead of humans to make traffic decisions. The ultimate goals are autonomous driving.

Vehicle detection based computer vision is an essential part of perception in autonomous driving. Research on visual thinking shows that visual perception accounts for more than 80% of human perception of the world [3]. How to introduce visual perception is a critical issue for autonomous driving.

However, designing a visual intelligence system that replaces substitute perception is a tricky task. As a critical part of intelligent perception systems, vehicle detection is one of the classic scientific issues. Thus, current perception-based autopilot algorithms are primarily solving the problem of assisting human drivers to ensure traffic safety.

The detection and tracking of vehicles ahead is a hot topic in the field of safety assisted driving, and it is an essential content in obstacle detection research. In recent years, many algorithms and implementation methods have been proposed at home and abroad, such as the use of vehicle linear geometric feature information, vehicle edge symmetry [4] or the use of specialized hardware (color CCD [5-7] and binocular [8-10] computer vision methods. Besides, there are optical flow-based

methods, template matching methods, training methods using neural networks [6], and multi-sensor information fusion methods [6,11].

This paper concluded vehicle detection by feature engineering methods: manual feature design and data-driven feature engineering. Manual feature design is the typical feature extraction algorithms before deep learning. The performance and speed of such algorithms are lower than deep learning in the general category and most scenarios, however, the feature extraction design method ideas therein are still influencing the design and implementation of object detection, vehicle detection algorithms. There are specific application scenarios where these algorithms still have an intense life in the vehicle inspection task. Data-driven feature engineering methods are the algorithms for proposing features based on deep learning. These features are characterized by distributed representation, rich in semantic information. Unlike manual feature design, this approach directly uses neural networks for self-learning-type extraction of features of images. As a result, deep learning algorithms are gradually becoming more and more popular among researchers and have become the best performing algorithms in object detection and vehicle detection.

Also, this paper summarizes the 3D vehicle detection algorithms based on stereo perception.3D vehicle detection provides the vehicle detection algorithm with more perceptual information related to autopilot.

2. Classification based on feature extraction

Existing image-based vehicle detection methods can be divided into two main categories according to their feature extraction techniques, manual feature design, and data-driven characterization learning.

2.1. vehicle detection algorithm based on manual feature design

Vehicles have very distinctive features, textures in the study of salient image features. This nature makes it possible to conduct vehicle detection based on pattern recognition methods. In order to learn these features, the researchers relied on several feature extraction methods for feature extraction. Among them, the histogram of oriented gradients (HOG) feature [12], Haar feature [13], and local binary pattern (LBP) are standard extraction operators. These three features are not specifically designed for vehicle detection, but based on the vehicle's image features and statistical method learning, the study shows that these features apply to the vehicle detection algorithm.

Hu [14] proposed the multi-resolution directional gradient histogram (HOG) feature and local binary pattern (LBP) fusion algorithm. The training samples are enlarged and reduced to form three different resolutions, and the HOG features are extracted respectively and superimposed to extract the LBP features of the training samples. The superimposed HOG features and LBP features are fused to obtain a feature vector. Compared with a single feature vehicle detection method, the detection rate is significantly improved. This method uses the histogram intersection kernel support vector machine to train and classify the features, which has the advantages of fast classification speed and high efficiency.

Wei [15] proposed a vehicle detection algorithm combining the Haar and HOG Features. They got the prospect region of interest (ROI) extracted by the Haar features, then use the inherent advantages of HOG features to detect target vehicles. Moreover, the HOG features of the ROI areas selected by the cascade structured AdaBoost algorithm, then a trained support vector machine (SVM) extracted the more precise results. Experimental results show the algorithm performs well in real-world scenarios, and ensure the detecting accuracy and time efficiency.

Huang [16] presents a vehicle detection and inter-vehicle distance estimation algorithm. Their work performs well on urban/suburban roads. In their work, vehicle features extracted by HOG features, and detect vehicle area with SVM classifier. Amazingly, the algorithm detects vehicle-based on vehicle shadow at the bottoms of vehicles.

The machine learning method based on SVM or neural network is computationally intensive and time-consuming in vehicle detection, and the detection performance needs to be further improved. Based on the Haar feature [13] and the AdaBoost classifier [17], it combines moving object detection

methods to identify vehicles in surveillance videos automatically, and proposes two improved algorithms:

(1) BD_HaBoost is based on the AdaBoost classification and uses the target object difference method for secondary recognition, which improves accuracy;

(2) TD_HaBoost uses the background difference method to extract the moving area to eliminate the influence of interfering objects in the background. Then it is classified by the AdaBoost classifier, which improves the detection rate and shortens the detection time.

Three videos were used in the experiment. The first video was a surveillance video on a straight road, the second video was a surveillance video of a traffic intersection, and the third video was a video outside a large shopping center.

The two algorithms are compared with the Haar + BP neural network background difference algorithm (HaBP) and the HOG + SVM + background difference algorithm (HoS-VM). The experimental results are shown in Table 1.

Table 1. Comparison of the detection time of each algorithm

Algorithm	Single frame average detection time / ms		
	Video1	Video2	Video3
HaBP	33.72	33.48	32.33
HoSVM	145.50	145.42	143.20
TD_HaBoost	65.55	47.70	45.14
BD_HaBoost	30.78	29.88	21.23

2.2. Data-driven based vehicle detection algorithms

Compared with the feature engineering methods, the deep learning object detection algorithms is the data-driven algorithm for feature extraction and object detection. The deep learning methods extract features from the statistic of image data and automatically learn the appearance features of objects. Moreover, the convolutional neural networks (CNN) features have more representative characteristic, and the feature learning process is similar to for human visual mechanism [18]. Due to the different structure of networks and the feature learning progress, the deep learning object detection algorithms are categorized in two directions, two-stage object detection, and one-stage object detection.

2.2.1. Two-stage detection algorithms.

The two-stage detection algorithm divides the detection process into two stages, generating candidate regions (region proposals), and classifying the candidate regions (generally with location refinements).

R-CNN [19] is the classical algorithm in object detection, and it extracted vehicle features by Convolutional Neural Networks (CNN). The method concludes the detection task as two-stage. Firstly, the selective search algorithm to generate proposal regions of interest (ROI). Secondly, use the SVM classifier to detect the most accurate region and the object's type.

Fast RCNN [20] is the upgraded algorithm of R-CNN, and its innovations are the ROI pooling layer, which solves the duplicate computation in R-CNN. Moreover, Fast R-CNN used a softmax layer instead of an SVM classifier in R-CNN to improve the detection results.

Faster RCNN [21] is the upgraded algorithm of Fast R-CNN, and the algorithm used the RPN networks instead of selective search methods in R-CNN and Fast RCNN. Smooth L1 loss is another innovation of Faster RCNN; the loss function smoothes the training progress and increases the detection results.

Region-based fully convolutional network (R-FCN) [22] attempts to improve detection performance based on Faster RCNN and FCN and the main contributions are a. introduction of FCN to achieve more network parameters and feature sharing (compared to Faster RCNN) b. solving the problem of location sensitivity deficiencies of fully convolutional networks (using position-sensitive score maps).

In conclusion, the two-stage detection algorithms have precise detection results in vehicle detection. However, its inference time can't achieve real-time.

2.2.2. One-stage detection algorithms.

The main idea of one-stage detection is to conduct intensive sampling in different locations of the images uniformly, sampling can use different scales and aspect ratios, and then use CNN to extract features and then directly classify and regress, the whole process requires only one step. Hence, its advantage is fast, but a critical disadvantage of uniform intensive sampling is that training is more complicated, resulting in slightly lower model accuracy.

SSD [23] transform the inspection task into a uniform, end-to-end regression problem, and obtain the position and classification simultaneously by only one process. The idea of transforming detection into regression was inherited from YOLO, and a similar Prior box was proposed based on Anchor in Faster RCNN to locate and classify targets at once; a Pyramidal Feature Hierarchy-based detection approach was added, i.e., predicting targets on a feature map of different sensory fields.

YOLO [24-26] uses the CNN network to implement the detection, the training and prediction process is end-to-end, the algorithm is fast and straightforward, YOLO does the convolutional calculation of the whole picture, so it has the advantage of a larger field of view during the detection, and is not easy to misjudge the background. The full convolutional layer serves the purpose of the attention module. Besides, the generalization capability of YOLO is well, and the model robustness is high when migrating. Moreover, the new feature extraction network (Darknet-19), adaptive anchor box, and multi-scale training make YOLO detection performance well.

RetinaNet [27] solves this problem that the loss of simple samples can cover the loss of a large number of complicated cases, and it is a One Stage algorithm model with an accuracy comparable to the Two-Stage detection algorithm. However, compared with YOLO, SSD, RetinaNet's inference time is slow.

In order to better understand the vehicle detection aspects of the five models, the experiments were divided into three scenarios using vehicle images from the public data set KITTI as shown in Table 2. And get three scenarios under the P-R curve as shown in Figure 1.

Table 2. Performance of five deep models for vehicle detection on KITTI dataset

Method	Average Precision (%)			FPS	Total Number
	Easy	Moderate	Hard		
Faster R-CNN	81.09	57.47	48.37	0.68	7518
R-FCN	81.24	79.49	66.01	0.31	7518
SSD	56.63	45.93	38.91	14.15	7518
YOLOv3	58.56	45.98	38.23	8.17	7518
RetinaNet	89.93	78.85	68.73	3.59	7518

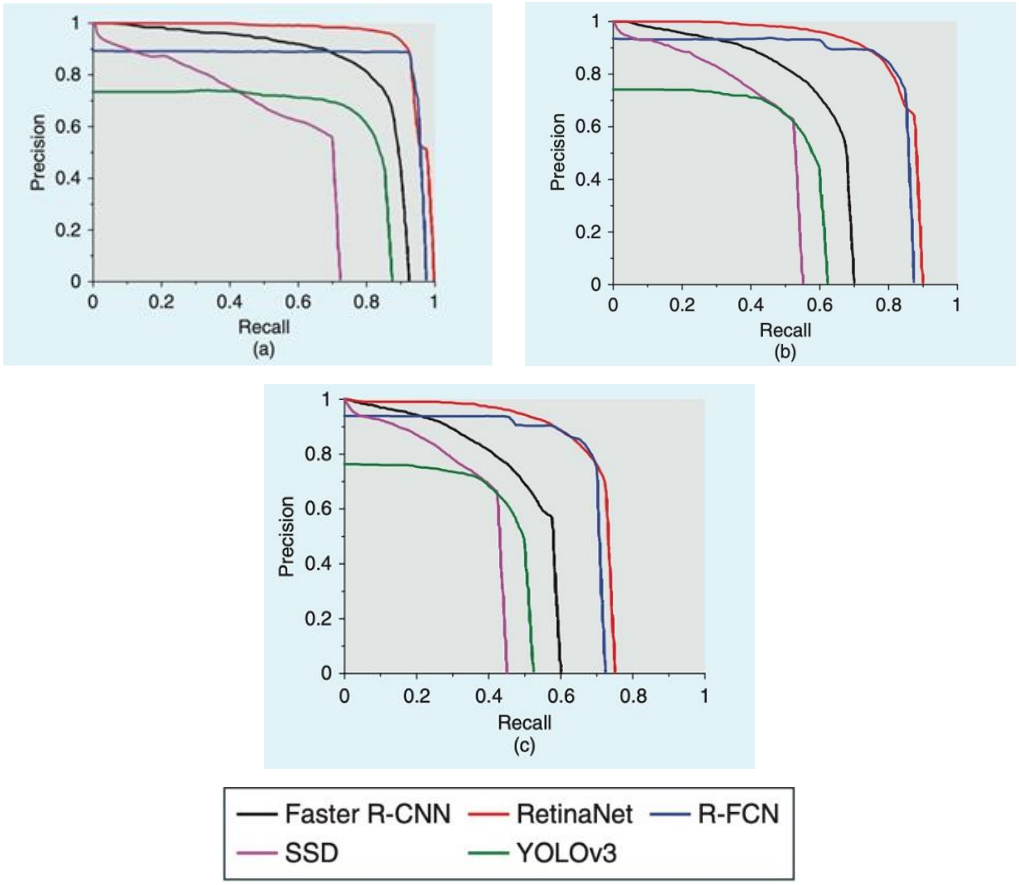

Figure 1. The PR curve of five deep models for vehicle detection. (a) Easy object, (b) moderate object, and (c) hard object.

3. Stereo perception

With the emergence of Faster-RCNN, 2D object detection has reached an unprecedented boom, and various new methods are emerging, and hundreds of thoughts are competing. However, in the application scenarios of drones, robots, and augmented reality, ordinary 2D detection cannot provide all the information needed to sense the environment.

At present, 3D target detection is in a period of rapid development, and at present, it is mainly using the monocular camera, binocular camera, multi-line LIDAR to carry out 3D target detection. However, with the continued industrialization of LiDAR, the cost is decreasing, and there are now some technical solutions for the combined use of LiDAR with a single-eye camera and a small number of lines.

YOLO 3D [28], an extensive multi-task 3D object detection network applied to the Apollo framework, performs lane line identification and object detection with stereo information. The Encoder Module is Yolo's Darknet, which adds a deeper convolutional layer to the original Darknet and an anti-convolutional layer to capture more valuable image context information. The algorithm uses high-resolution multi-channel feature maps, which capture image detail and provide in-depth, low-resolution multi-channel feature maps that provide the encoder with more image context information. The Decoder is divided into two parts, one is semantic segmentation for lane line detection, and the other is object detection, the vehicle detection part is based on YOLO, and also outputs 3D information such as the direction of the vehicle.

AVOD [29] algorithm using visual and radar information, its input RGB image, and BEV (Bird Eye View). The algorithm uses the FPN network to obtain the full resolution feature map, extracts the

corresponding regions of the two feature maps for fusion, and selects a 3D proposal for 3D vehicle detection.

Although there are many pieces of research focused on 3D object detection, many problems still exist in the 3D vehicle detection.

1. The robustness of the object blocking, truncation, the surrounding dynamic environment.

2. The detection algorithm mainly relies on the surface texture or structural features of the object, which can easily confuse the detection results.

3. The algorithm efficiency is not high.

Because a 3D bounding box is the smallest rectangle that surrounds a target object in the real 3D world, a 3D bounding box theoretically has 9 degrees of freedom, 3 for the position, 3 for rotation, and 3 for dimensional size. For the autopilot scenario, most of the objects are placed horizontally on the ground, so by assuming that the objects are placed horizontally on the ground, you can set the rolling and tilting angle to zero relatives to the horizontal plane. In contrast, the bottom plane is part of the horizontal plane, so that 3 degrees of freedom can be omitted, and 6 degrees of freedom, so 3D target detection is also a target object 6D pose prediction problem.

4. Conclusion

Vehicle detection is a computer that determines whether there is a vehicle in a given image and video and determines the location of the vehicle. Vehicle detection is the basis and premise for researches such as parking lot management, vehicle tracking, and vehicle license plate recognition. But in the actual operation, there are still many difficulties to be solved in vehicle detection, such as occlusion, lighting, and object shape changes. This paper introduces some technologies of vehicle detection from the two directions of deep learning and vision, and compares the real-time and correctness of different algorithms.

Although the current vehicle detection technology is relatively mature, it still needs to be improved in small object, occlusion and real-time. In terms of vehicle detection technology, how to use the existing technology to detect small object vehicles is still the focus of research in various fields. With the development of basic sciences of artificial intelligence and visual computing theory, as well as the improvement of sensor technology and the continuous improvement of the cost performance of computers, vision-based vehicle detection and tracking technology will make future cars develop more intelligent and practical.

Acknowledgments

Ying Zheng has contributed to this paper by organizing the materials and literature research. Chaochao Meng thanks Yan Ma for her patience and understanding.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 61932012) and National Natural Science Foundation of China (Grant No. 91420202).

References

[1] Bank, T. W. (2018). Global Road Safety Facility (GRSF) annual report 2017. http://documents.shihang.org/curated/zh/677161516192289928/Global-Road-Safety-Facility-GRSF-annual-report-2017.

[2] Kapp C. WHO acts on road safety to reverse accident trends [J]. The Lancet, 2003, 362(9390): 1125-1125.

[3] Carrascal S, Magro M, Anguita J M. Acquisition of Competences for Sustainable Development through Visual Thinking. A Study in Rural Schools in Mixco, Guatemala [J]. Sustainability, 2019, 11(8).

[4] Bensrhair A, Bertozzi M, Broggi A. A cooperative approach to vision-based vehicle detection[C]. IEEE intelligent transportation systems, 2001: 207-212.

[5] Heisele B. Motion-based Object Detection and Tracking in Color Image Sequence[C]. asian conference on computer vision, 2000..
[6] Steux B, Laurgeau C, Salesse L. Fade: a vehicle detection and tracking system featuring monocular color vision and radar data fusion [J]. Information Visualization, 2002: 632-639.
[7] Betke M, Haritaoglu E, Davis L S. Real-time multiple vehicle detection and tracking from a moving vehicle [J]. Machine Vision & Applications, 2000, 12(2):69-83.
[8] Bensrhair A, Bertozzi A, Broggi A. Stereo vision-based feature extraction for vehicle detection[C]// Intelligent Vehicle Symposium. IEEE, 2002:465-470 vol.2.
[9] Labayrade R, Aubert D. In-vehicle obstacles detection and characterization by stereovision [J].in 'Proceedings the 1st International Workshop on In-Vehicle Cognitive Computer Vision Systems, 2003:13--19.
[10] Labayrade R, Aubert D, Tarel J. Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation[J]. Information Visualization, 2002: 646-651.
[11] Wang C, Thorpe C, Suppe A. LADAR-based detection and tracking of moving objects from a ground vehicle at high speeds[C]. Intelligent vehicles symposium, 2003: 416-421.
[12] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. Computer vision and pattern recognition, 2005: 886-893.
[13] Mita T, Kaneko T, Hori O. Joint Haar-like features for face detection[C]. International conference on computer vision, 2005: 1619-1626.
[14] HU Qing-xin, JIAO Wei, GU Ai-hua. Vehicle detection method based on multi-features fusion and intersection kernel SVM [J]. Journal of hefei university of technology (natural science), 2016.
[15] Wei Y, Tian Q, Guo J. Multi-vehicle detection algorithm through combining Harr and HOG features [J]. Mathematics and Computers in Simulation, 2018: 130-145.
[16] Huang D, Chen C, Chen T. Vehicle detection and inter-vehicle distance estimation using single-lens video camera on urban/suburb roads ☆[J]. Journal of Visual Communication and Image Representation, 2017: 250-259.
[17] Viola P A, Jones M. Robust real-time face detection[C]. International conference on computer vision, 2001, 57(2): 137-154.
[18] Lecun Y, Bengio Y, Hinton G E. Deep learning [J]. Nature, 2015, 521(7553): 436-444.
[19] Jia Y, Shelhamer E, Donahue J. Caffe: Convolutional Architecture for Fast Feature Embedding[C]. acm multimedia, 2014: 675-678.
[20] Girshick R. Fast R-CNN[C]. International conference on computer vision, 2015: 1440-1448.
[21] Ren S, He K, Girshick R. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.
[22] Dai J, Li Y, He K. R-FCN: object detection via region-based fully convolutional networks[C]. Neural information processing systems, 2016: 379-387.
[23] Liu W, Anguelov D, Erhan D. SSD: Single Shot MultiBox Detector[C]. european conference on computer vision, 2016: 21-37.
[24] Redmon J, Divvala S K, Girshick R. You Only Look Once: Unified, Real-Time Object Detection[C]. Computer vision and pattern recognition, 2016: 779-788.
[25] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[C]. Computer vision and pattern recognition, 2017: 6517-6525.
[26] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement [J]. arXiv: Computer Vision and Pattern Recognition, 2018.
[27] Lin T, Goyal P, Girshick R. Focal Loss for Dense Object Detection[C]. International conference on computer vision, 2017: 2999-3007.
[28] Mousavian A, Anguelov D, Flynn J. 3D Bounding Box Estimation Using Deep Learning and Geometry[C]. Computer vision and pattern recognition, 2017: 5632-5640.

[29] Ku J, Mozifian M, Lee J. Joint 3D Proposal Generation and Object Detection from View Aggregation[C]. Intelligent robots and systems, 2018: 1-8.