Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230306

Review of object detection challenges in autonomous driving

Shenxuan Cao

North Cross School Shanghai, Building 2, Lane 803, Shuangcheng RD, Shanghai, CHN, Independent 200940

28253543@qq.com

Abstract. This paper presents a comprehensive review of object detection in autonomous driving applications. The classical object detection network is presented, along with several well-known online resources and benchmark methods. A thorough review of the challenges in object detection for autonomous driving is provided, along with potential solutions to these challenges. By exploring the current state of object detection in autonomous vehicles, this paper aims to contribute to the ongoing efforts to improve the safety and efficiency of autonomous driving technology.

Keywords: auto-driving, object detection, deep learning, machine learning.

1. Introduction

Autonomous driving brings many benefits. Firstly, it can reduce the occurrence of accidents because artificial intelligence is more likely to follow traffic rules and can predict pedestrians' next moves, thereby reducing the accident rate. Compared to experienced human drivers, autonomous driving vehicles have faster and more reliable AI prediction capabilities and braking response speeds. Secondly, autonomous driving vehicles can save more lives. According to statistics, 90% of road traffic accidents are caused by human negligence or error, including factors such as distraction, fatigue, drunk driving, and speeding. Globally, 1.3 million people die from car accidents every year, with China alone accounting for over 60,000. Autonomous driving vehicles will eliminate accidents caused by human negligence and errors, thus greatly reducing the accident rate.

The core of autonomous driving lies in object recognition [1,2]. Only by correctly recognizing road objects and conditions can effective intelligent judgments be made. Existing autonomous driving object recognition can be divided into three categories: radar-based object recognition [3,4], camera-based object recognition, and fusion-based object recognition [5,6,7].

The radar-based object recognition solution comes from the invention of laser radar sensors, which were inspired by bats. Laser radar is a technology that is widely used in modern society, involving science, military, meteorology, transportation, and other fields. In autonomous driving technology, laser radar serves as a sensor that can obtain information about the surroundings of the vehicle by emitting laser beams. It has the advantages of high resolution, long detection distance, and good concealment. However, laser radar loses its effectiveness in harsh weather conditions such as rain and snow, and it is relatively expensive.

Therefore, some automakers have adopted a camera-based object recognition approach. This approach uses cameras as the "eyes" of the vehicle to collect information and simulate the driving

DOI: 10.54254/2755-2721/8/20230306

activities of human drivers. For a long time, Tesla has insisted on using cameras and chip-based neural networks to achieve autonomous driving technology. This method has a lower cost and can process collected image data through a perceptual neural network architecture to construct a three-dimensional vector space of the real world. Tesla vehicles have been equipped with 8 cameras, with an average price of around $10, which is approximately $600 (approximately 4,000 yuan) cheaper than laser radar. However, the camera-based approach has limitations in practical applications, such as night driving.

To achieve higher safety, some autonomous driving systems adopt an integrated approach that combines the advantages of both radar and cameras. During driving, the autonomous driving system collects information from both radar and camera sensors, and verifies the reliability of the information through mutual validation. When one sensor is unable to effectively collect information, the system can rely on the other sensor to provide additional safety assurance.

This paper focuses on the field of object recognition in camera surveillance. Early research on object recognition was based on classical frameworks for feature extraction and classification [8]. Since computers can only recognize numbers, we need to convert images into numbers to enable the computer to "see". To achieve this goal, we need to extract important and useful features from the image and train the computer to learn how to extract and recognize these features, which is the process of feature extraction. Based on the extracted feature information, the general principle of image classification and recognition is to distinguish an image from other different categories of images in order to identify the category to which it belongs [9,10].

Feature extraction can be divided into various methods such as basic statistical features, gray level co-occurrence matrix, feature dimension reduction, local binary pattern, and pedestrian detection HOG+SVM. Common basic statistical features include region descriptors such as perimeter, area, and mean, as well as histograms and gray level co-occurrence matrices. Features are characteristics or attributes used to differentiate one object from another, such as brightness, edges, textures, and colors. Gray level co-occurrence matrices are used to describe the relationship between two pixels in terms of gray levels, which helps to extract texture features from the image. However, due to the relatively large computation required for gray level co-occurrence matrices, it is generally necessary to compress the gray levels of the image to reduce the size of the matrix. Feature dimension reduction refers to the transformation of original features into a new feature space to reduce the dimensionality of the features. When performing feature dimension reduction, it is necessary to retain the main information and avoid excessive compression that leads to a decrease in classifier performance. Local binary pattern is a commonly used image texture feature descriptor that is simple and insensitive to gray level changes, and is widely used in the field of computer vision. In pedestrian detection, HOG features of positive and negative samples are generally used for training to obtain the SVM classifier model. Then, this model is used to generate detection sub-windows, which are tested on negative samples to obtain Hard Examples. Finally, the HOG features of Hard Examples are extracted and combined with the features of the first step for training to obtain the final pedestrian detection sub-windows. When selecting features, we hope to choose those features that have small differences among the same class of images, but large differences among images of different categories (i.e., large inter-class distance), which are called discriminative features.

Thanks to the significant development of GPUs, neural networks are now the basis for classification recognition methods. Various types of neural networks have been proposed to break existing recognition accuracy records, from early models like AlexNet, VGG, U-Net, and ResNet to the currently popular transformer. Increasingly large networks can be trained and validated, ensuring that classification recognition based on deep learning networks becomes increasingly accurate.

Based on existing research on deep learning networks, this paper provides a comprehensive review that includes a summary of classical methods, an introduction to available resources, and an overview of evaluation criteria. In addition, this paper will discuss the challenges in practical applications, including both human and natural factors.

The structure of this paper is as follows. The second paragraph will introduce classical methods and the latest research. The third paragraph will introduce existing toolkits, including datasets and evaluation

DOI: 10.54254/2755-2721/8/20230306

criteria. The fourth paragraph will discuss the challenges in practical applications. The fifth paragraph will conclude the paper.

2. Introduction to classical deep neural networks

In this section, this paper will introduce classical neural networks, including three classic object detection networks: R-CNN [11], Fast R-CNN [12], and Faster R-CNN [13]. These three networks are among the most famous object detection networks in the field of deep learning, and their emergence has not only greatly promoted the development of object detection but also had a profound impact on the field of computer vision.

2.1. R-CNN

R-CNN was the first end-to-end object detection framework proposed by Ross Girshick and colleagues in 2014. R-CNN (Region-based Convolutional Neural Network) is a region-based convolutional neural network that can classify and locate candidate object regions in an image. The workflow of R-CNN consists of three steps: first, using the selective search algorithm to extract candidate object regions; then, performing convolution and feature extraction on each candidate region and inputting the features into a support vector machine (SVM) for classification; finally, using a regressor to refine the position of the object. Although R-CNN has high detection accuracy, its speed is very slow and cannot be applied to real-time object detection.

2.2. Fast R-CNN

Fast R-CNN is an object detection network proposed by Ross Girshick and colleagues in 2015, which is an improvement over R-CNN. Fast R-CNN integrates the selective search step into the network, enabling simultaneous extraction of regions and features. These features are then input into a RoI pooling layer to obtain a fixed-length feature vector. This feature vector can be input into a fully connected layer for classification and localization. The advantages of Fast R-CNN are end-to-end training and faster detection speed with higher accuracy.

2.3. Faster R-CNN

Faster R-CNN was proposed in 2015 by Shaoqing Ren et al. as an improved object detection network based on Fast R-CNN. Faster R-CNN introduces a subnetwork called the Region Proposal Network (RPN) which can directly extract candidate object regions from images without requiring selective search algorithms. The RPN can simultaneously generate object candidate regions and corresponding bounding box regression values, which are then inputted into Fast R-CNN for classification and localization. The advantages of Faster R-CNN include faster speed, higher accuracy, and the ability to train the entire network directly. As one of the most popular object detection networks, Faster R-CNN has been widely applied in various aspects of computer vision.

3. Methods for online resource and benchmarking

In this section, we will present various methods for online resource utilization and benchmarking, which will include online training and evaluation techniques.

3.1. Benchmarking methods

Traditional benchmarking methods include Mean Average Precision, Intersection over Union (IoU), Precision and Recall, and F1 score.

1. Average Precision (AP): Average Precision is a widely used evaluation metric in object recognition algorithms. It measures the area under the precision-recall curve and reflects the overall accuracy and completeness of object detection results. The formula for Average Precision is as follows:

$$AP = \int_{[0,1]} p(r)dr$$

where p(r) is the maximum precision value when the recall rate is r, that is:

DOI: 10.54254/2755-2721/8/20230306

$$p(r) = max\{p \mid r' \geq r\}$$

Here, p represents precision, r represents recall rate, and r' represents the number of positive samples detected by the detector in all positive samples in the dataset.

It should be noted that the Average Precision value is usually calculated at different Intersection over Union thresholds (IoU thresholds). That is, for each IoU threshold, an AP value is calculated. The final Average Precision value is the average of these AP values.

2. Intersection over Union (IoU): IoU is a commonly used metric for evaluating the accuracy of object detection and localization. It measures the degree of overlap between the predicted bounding box and the true bounding box. The formula for IoU is as follows:

$$IoU = Intersection(A,B) / Union(A,B)$$

Here, A and B represent the two bounding boxes, Intersection (A, B) represents their intersection area, and Union (A, B) represents their union area. The value of IoU ranges from 0 to 1, and a higher value indicates a greater degree of overlap between the two bounding boxes. Generally, a predicted box is only considered a correct detection result when its IoU value is greater than a certain threshold. IoU is often used to calculate loss functions (such as GIoU loss in YOLOv5), helping object detectors learn more accurate detection boxes.

3. Precision and Recall: Precision and recall are fundamental measures for evaluating object detection algorithms. Precision measures the proportion of correctly detected objects among all detections, while recall measures the proportion of correctly detected objects among all ground truth objects. The formulas for precision and recall are as follows:

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

Here, TP represents true positive, which is the number of samples correctly predicted as positive, FP represents false positive, which is the number of negative samples incorrectly predicted as positive, and FN represents false negative, which is the number of positive samples incorrectly predicted as negative.

It is important to note that precision and recall often have a trade-off. Increasing the precision of the classifier usually means reducing the number of false positives, but may increase the number of false negatives, thereby reducing recall. On the other hand, improving the recall of the classifier means reducing the number of false negatives, but may increase the number of false positives, thereby reducing precision. To comprehensively evaluate the performance of a classifier, the F1 score is often used to balance precision and recall. Its formula is:

$$F1 = 2 * (Precision * Recall) / (Precision + Recall)$$

4. F1 Score: The F1 score is a metric that combines precision and recall into a single performance measure for object detection algorithms. It is the harmonic mean of precision and recall, reflecting the completeness and correctness of object detection results. The formula for F1 score is:

$$F1 = 2 * (Precision * Recall) / (Precision + Recall)$$

Here, Precision represents the proportion of positive samples predicted by the classifier that are actually positive, while Recall represents the proportion of positive samples that are correctly predicted by the classifier. The F1 score ranges from 0 to 1, with a higher score indicating better classifier performance. Generally, the classifier's performance reaches its optimal level when the F1 score reaches its maximum.

It is important to note that the F1 score may not be a good evaluation metric for imbalanced datasets, as in this case, the classifier's performance may be stuck at a low level and unable to further improve by increasing either recall or precision. Therefore, in practical applications, other evaluation metrics need to be considered to comprehensively evaluate classifier performance. In the field of object detection, the F1 score is usually used together with precision and recall to evaluate the performance of object detectors.

DOI: 10.54254/2755-2721/8/20230306

In practice, the F1 score can help quickly compare the performance of different object detectors and find the optimal one.

3.2. Online dataset

The introduced online dataset including PASCAL VOC and COCO.

PASCAL VOC (Visual Object Classes) is a widely-used dataset in object recognition algorithms. It is a commonly used database in the fields of object detection and image classification, jointly developed by the University of Oxford in the United Kingdom and the Image Understanding Group in Germany [14]. The database contains 20 different categories of objects, including animals, vehicles, and people. Each category has about 100 images of different sizes and angles. The annotation method of the PASCAL VOC dataset adopts the "bounding box + category label" method, which annotates the position and category information of each object in the image together. The release and updates of the PASCAL VOC database have also sparked a lot of academic research and have become one of the benchmark datasets in the fields of object detection and image classification. The authors of PASCAL VOC include the University of Oxford in the United Kingdom, the Image Understanding Group in Germany, and others.

The COCO (Common Objects in Context) database is a larger and more challenging dataset than PASCAL VOC [15]. It contains over 330,000 images and covers 80 different categories of objects. Unlike PASCAL VOC, COCO not only annotates the bounding box and category of the objects but also the keypoint information of each object. The annotation method of COCO is more detailed, which also means that the dataset is more challenging. The release and updates of the COCO database have driven advances in the fields of object detection and image segmentation. The authors of COCO include institutions such as Microsoft and Cornell University.

These databases contain a large amount of data. The PASCAL VOC database includes over 11,000 annotated images of objects, while the COCO database includes over 330,000 images and 2.5 million annotated objects. They also provide average precision evaluation at different IoU thresholds and other performance metrics. These evaluation metrics provide standardized methods for the research and evaluation of object detection algorithms.

4. Analysis of practical challenges

Object recognition technology has wide applications in various fields, such as autonomous driving, intelligent security, unmanned aerial vehicles, medical imaging, and so on. However, despite the development of object recognition technology, there are still some engineering challenges that need to be addressed. This paragraph will list four engineering challenges in the field of object recognition and explore possible solutions.

1. Multi-object recognition and occlusion problems

In practical applications, it is often necessary to recognize multiple objects, and these objects may occlude each other. For example, in the autonomous driving scene, the driver needs to identify multiple vehicles, pedestrians, traffic lights, and so on, and these objects may occlude each other, leading to a decrease in recognition accuracy. To solve this problem, many current studies focus on multi-object tracking technology. This technology can simultaneously track multiple objects and predict their motion trajectories, thus reducing occlusion problems. Additionally, deep learning models, such as Mask R-CNN, can be used to handle occlusion issues.

2. Imbalanced dataset problem

The dataset in the field of object recognition often exhibits a class imbalance problem, where some classes have much more samples than others. For example, in face recognition datasets, some people's samples may have several times or even tens of times more samples than others, resulting in poor performance of the model on minority classes. To solve this problem, researchers have found that data augmentation techniques can be used to balance the dataset. For example, in image classification tasks, techniques such as random cropping, rotation, and flipping can be used to increase the number of training samples, thereby reducing the imbalanced dataset problem. Additionally, few-shot learning, meta-

DOI: 10.54254/2755-2721/8/20230306

learning, and other techniques can be used to address this issue.

3. Diverse object appearances and shapes

The appearance and shape of objects may change due to lighting, angle, occlusion, and other reasons, making it difficult for the model to accurately recognize them. For example, in face recognition tasks, photos of the same person may have significant appearance differences due to different shooting angles, lighting conditions, and other reasons. To solve this problem, data augmentation techniques can be used to simulate different appearance and shape changes. For example, in face recognition tasks, techniques such as image rotation, cropping, and noise addition can be used to generate diverse face images. Additionally, feature alignment techniques can be used to normalize the shape of objects, thereby reducing the impact of shape changes on recognition results.

4. Real-time and efficiency issues

In some real-time application scenarios, object recognition tasks need to be completed within a short time. For example, in the autonomous driving scene, objects on the road need to be recognized and corresponding decisions need to be made within tens of milliseconds. Therefore, the system approaches used in practical applications improve real-time and efficiency by using lightweight models and hardware accelerators. For example, lightweight models such as MobileNet and EfficientNet can be used to reduce the computational complexity and number of parameters of the model. Additionally, hardware accelerators such as GPU, FPGA, and ASIC can be used to improve the computational speed and efficiency of the model.

In summary, the engineering challenges in the field of object recognition mainly include multi-object recognition and occlusion problems, imbalanced dataset problems, diverse object appearances and shapes, and real-time and efficiency issues. The solutions to these problems include the use of multi-object tracking technology, data augmentation techniques, feature alignment techniques, lightweight models, and hardware accelerators. In the future, with the development of deep learning and computer hardware, object recognition technology will continue to advance and achieve new breakthroughs in various applications.

5. Conclusion

This academic manuscript presents a comprehensive overview of three prominent object detection networks, namely R-CNN, fast R-CNN, and faster R-CNN. In addition, the paper delves into several widely-recognized benchmark techniques utilized for assessing the efficacy of object detection methods. Furthermore, the study provides an insight into two well-established online datasets, PASCAL VOC and COCO, which are extensively utilized for training, validating, and testing object detection models. Crucially, the paper meticulously scrutinizes four practical challenges commonly faced by object detection systems and presents potential solutions to address them.

References

[1] Zou, Zhengxia, et al. "Object detection in 20 years: A survey." Proceedings of the IEEE (2023).

[2] Zaidi, Syed Sahil Abbas, et al. "A survey of modern deep learning based object detection models." Digital Signal Processing (2022): 103514.

[3] Niederlöhner, Daniel, et al. "Self-supervised velocity estimation for automotive radar object detection networks." 2022 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2022.

[4] Gonzales-Martínez, Rosa, et al. "Hyperparameters tuning of faster R-CNN deep learning transfer for persistent object detection in radar images." IEEE Latin America Transactions 20.4 (2022): 677-685.

[5] Wei, Zhiqing, et al. "Mmwave radar and vision fusion for object detection in autonomous driving: A review." Sensors 22.7 (2022): 2542.

[6] Zhou, Taohua, et al. "Bridging the view disparity between radar and camera features for multi-modal fusion 3d object detection." IEEE Transactions on Intelligent Vehicles (2023).

[7] Hwang, Jyh-Jing, et al. "CramNet: Camera-Radar Fusion with Ray-Constrained Cross-Attention for Robust 3D Object Detection." Computer Vision–ECCV 2022: 17th European Conference,

DOI: 10.54254/2755-2721/8/20230306

Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII. Cham: Springer Nature Switzerland, 2022.

[8] Diwan, Tausif, G. Anirudh, and Jitendra V. Tembhurne. "Object detection using YOLO: Challenges, architectural successors, datasets and applications." Multimedia Tools and Applications (2022): 1-33.

[9] Wang, Yi, et al. "Remote sensing image super-resolution and object detection: Benchmark and state of the art." Expert Systems with Applications (2022): 116793.

[10] Liang, Tingting, et al. "Cbnet: A composite backbone network architecture for object detection." IEEE Transactions on Image Processing 31 (2022): 6893-6906.

[11] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[12] Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE international conference on computer vision. 2015.

[13] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems 28 (2015).

[14] Vicente, Sara, et al. "Reconstructing pascal voc." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[15] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014.