Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Exploring Object Detection: Datasets, Metrics and Algorithms

Yifei Zhao *

Department of Data Sciences, School of Information, Beijing Forest University, Beijing, China

* Corresponding Author Email: zhaoyifei@bjfu.edu.cn

Abstract. Object detection, as one of the basic issues in the computer vision (CV) field, has huge application value in realm such as autonomous driving and face recognition. Since the object differences in images in different scenes, angles, and lighting environments are too great, the number and size of objects are also different, those greatly increase the difficulty of detection. Therefore, how to detect objects quickly and accurately has attracted widespread attention from researchers. This article provides a review from the perspectives of data and algorithms in accordance with the order of technological development. It mainly introduces commonly used data sets, evaluation indicators, traditional object detection algorithms, and object detection algorithms based on deep learning. Finally, this paper will discuss future research directions, explore what difficulties still exist in the field of target detection, and how researchers should further improve the performance and generalization capabilities of target detection to cope with more complex and changeable scenarios and needs.

Keywords: Object Detection; Datasets; CNN; Deep Learning; Computer Vision.

1. Introduction

Object detection, as a highly significant research direction in the CV field, has received widespread attention in recent years. In the past few years, with the speedy development of artificial intelligence (AI) technology, object detection has found widespread applications in various fields. Object detection, as its literal meaning suggests, is to identify specific objects in static or dynamic images. In the emerging field of autonomous driving, target detection plays an important role. It can identify pedestrians, vehicles and obstacles on complex roads so that the intelligent driving system can calculate and make correct driving decisions. In the medical field, target detection also performs well. It can accurately locate the location and type of lesions in medical images, assisting doctors in making more scientific and efficient diagnoses. In addition to autonomous driving and medical fields, object detection also performs well in fields such as agricultural pest identification and product quality testing in industrial production, providing convenience for people's production and life.

In object detection, the first task is to determine the target object's location, that is, a bounding box needs to be generated at the appropriate location. It is used to identify the accurate location of the object in the image. It is the basis for determining the object category and directly affects the subsequent target recognition accuracy. In actual applications, the same type of objects may have different sizes in the image, which requires the target detection algorithm to accurately detect objects at different scales. Target objects that are occluded and deformed in the image are also very important, which will make it difficult for general object detection algorithms to accurately identify them. The object detection algorithm also needs to consider the robustness of the algorithm when facing such problems to cope with complex scenes. In addition, in some scenarios, the target detection algorithm also needs to have high real-time performance, such as intelligent driving. Therefore, analyzing and processing images in a short time is also an important issue in object detection.

This paper mainly reviews the existing object detection technology in chronological order, and introduces the aspects of evaluation metrics, datasets, and common object detection algorithms. Among them, in terms of algorithms, it will introduce them in two stages: traditional target detection algorithms (before 2014) and deep learning-based target detection algorithms (after 2014).

2. Metrics and Datasets

2.1. Metrics

In order to study object detection in a better way, it's crucial to know the related indicators and datasets of those type of problems in advance. In terms of performance evaluation indicators, there usually have two different perspectives: accuracy and speed. Speed is well understood and intuitive. It represents the time it takes from the beginning of recognition to the recognition of the object. But the accuracy is kind of abstract that need some redefined variables to measure. In recent years, Average Precision (AP) [1] was the most frequently used evaluation for detection. Before studying the variables of AP, we need to know some relevant basic definitions first (those definitions are based on confusion matrix).

True Positive (TP). A positive sample is predicted by the classifier, and it matches the actual result.

False Positive (FP). A positive sample is predicted by the classifier, and it matches the wrong result.

True Negative (TN). A negative sample is predicted by the classifier, and it matches the actual result.

False Negative (FN). A positive sample is incorrectly predicted by the classifier, whereas the actual result is positive.

After knowing those basic definitions, we can easily calculate two concepts that are crucial for object detection: precision (P) and recall(R) which are defined as:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{1}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

Precision is the percentage of correct predictions among all prediction results, which represents the model's ability to identify relevant objects. Recall is the percentage of correctly predicted objects among all known objects, indicating the model's ability to find all relevant objects.

There is also a related concept called intersection over union (IoU), which can measure the object location accuracy. In object detection, IoU represents the proportion of the predicted bounding box's area (Bp) that intersects with the ground-true bounding box (Bgt) to the total area.

$$\mathrm{IoU} = \frac{\mathrm{area}(\mathrm{Bp} \cap \mathrm{Bgt})}{\mathrm{area}(\mathrm{Bp} \cup \mathrm{Bgt})} \tag{3}$$

In practical applications, we usually set a threshold for IoU, such as 0.5. When the actual IoU is greater than the threshold, the Bp will be categorized as "detected", or it will be categorized as "missed" [1]. So that will influence the calculating of precision and recall.

By calculating different P and R in different threshold, we can get many different sets of values of precision and recall. Putting them into the coordinate system and connect them together, and we get the precision-recall curve. What we called AP is the integral of the precision × recall curve (p-r curve), which is the graph's area enclosed by the p-r curve. The mean AP (mAP) is simply average the AP of all kinds of classes. It meanly used to measure the accuracy of models over each classes.

2.2. Datasets

In the training process of object detection, datasets with large amounts of data and small bias play an important role in training better models. There are so many well-known datasets, for instance,

PASCAL VOC [2], MS-COCO, Open Images, DOTA [3], etc. They target different fields, have different features, scales and usage.

PASCAL VOC07. The PASCAL Visual Object Classes (PASCAL VOC) [2] is a data set in CV field. It's one of the most important datasets in the early CV projects, which contains 20 classes commonly found in daily life, such as "car", "chair", "cat", "dog", "person", etc. Each image has its own annotation, which identifies the object bounding box' location and category of object in the image.

Except the dataset, PASCAL VOC also organizes the PASCAL VOC Challenge every year. It attracts a number of engineers and developers to participate. Therefore, it also benefits to the advancement of CV field and promote algorithms' effect in real-world scenarios.

DOTA. Dataset for Object detection in Aerial images (DOTA) [3] is a dataset for high-resolution Earth image taken by satellite. It contains images of many kinds of scenes and geographical locations, e.g. "Large-vehicle", "Bridge", "Ship", "Basketball court". Mostly, in this dataset, more than one objects are contained in images and they usually have different size, shape and direction. Images in the data set are annotated in detail, and the annotation information includes bounding box location, category label and occlusion of the objects. Those annotations allow developers to use the data set for supervised learning.

The DOTA dataset is usually used in the field of remote sensing to train models that can identify specific targets (such as cars, ships, etc.) in satellite images to achieve remote management of docks, bridges, airports, etc.

3. Object Detection Methods

In the past 20 years, under general knowledge, object detection has passed through two stages, named "traditional object detection stage (before 2014) "and "deep learning based object detection stage (after 2014) [4]." Below, we will summarize some important algorithms and detectors in these two periods.

3.1. Traditional Object Detection Algorithm

3.1.1. Classic Methods

Viola-Jones Detector. Viola-Jones Detector [5] is a traditional face detection system. It is mainly used to detect whether an image contains a face and the face's location in the image. At that time, it was known for its excellent speed and accuracy make it the algorithm of choice in many real-time applications. The Viola-Jones algorithm mainly detects objects through cascades. This algorithm first uses the Haar cascade method to extract features in the image, and calculates the differences between pixels to represent the features through a sliding window. Then uses the Adaboost algorithm [5] to select distinguishing features among all features that is, using a decision tree or other basic classification methods to form a series of weak classifiers, and then iteratively train the weak classifiers to construct a strong classifier. Finally, use the cascade method to combine the classifiers formed in the previous step to form a cascade classifier. The Viola-Jones detector can quickly exclude non-target areas in the image and locate the target accurately and quickly.

HOG Detector. Histogram of Oriented Gradients (HOG) [6] is a feature descriptor for object detection, proposed in 2005. Its main idea is to divide the image into small units, quantify the gradient direction within the unit into a certain direction interval, and then use the histogram of these direction intervals to describe the gradient characteristics of the image. The HOG algorithm first preprocesses the input image, such as grayscale and normalization. Then use filters such as Sobel mask [6] to calculate the gradient direction and intensity of each pixel in the image. The image is then divided into many cells, the direction and intensity of the gradient within each cell are calculated, and these gradients are statistically generated to generate a histogram that can describe the image. Finally, the generated image is normalized, and the histograms of all cells are combined into a vector, which is the final HOG feature descriptor. HOG features combined with SVM classifier [6, 7] have been

extensively used in image recognition, particularly in pedestrian detection, leading to significant success.

In most traditional object detection algorithms, engineers need to manually calculate features, such as Haar [5], HOG, etc. This requires engineers to have certain professional knowledge in the detection field and does not have sufficient generalization capabilities. Since traditional methods usually only consider local information in the image, the recognition accuracy is not high in complex or obstructed situations. In the Viola-Jones algorithm, although the calculation speed is relatively fast, it is easily affected in complex scenarios, resulting in reduced calculation efficiency. All above are common shortages in traditional object detection algorithms. Moreover, traditional methods usually divide tasks into multiple steps, such as feature extraction, classification, etc., which lack end-to-end learning and are less efficient.

3.2. Object Detection Algorithm based on Deep Learning

3.2.1. Application of Deep Learning in the Object Detection

Deep learning has greatly promoted the development of object detection. Among them, Convolutional Neural Network (CNN) is a revolutionary innovation, which provides new ideas for the field of object detection.

Among existing deep learning methods, there are mainly two object detection paradigms: single-stage detection and two-stage detection. Typically, a single-stage detector treats the object detection problem as a separate regression problem and achieves the purpose of identifying objects by predicting the positions and categories of many bounding boxes on the image. It has the characteristics of fast recognition speed and is suitable for real-time performance. Scenarios with high demand. The two-stage detector will first generate candidate areas and then classify the objects in the candidate areas. Therefore, compared with the single-stage detector, it can identify objects more accurately and is suitable for scenes with high accuracy requirements.

Deep learning also brings many benefits to object detection, for example, applying previously trained models to object detection through transfer learning, thereby skipping parts of training process improving training efficiency.

3.2.2. Classic Deep Learning Models

R-CNN Series. Region-CNN (R-CNN) [8], created by Ross Girshick, it marks the initial successful application of deep learning in object detection. It is a two-stage detector. In R-CNN [8], the Selective Search (SS) method is first utilized to segment the input image into 1k-2k region proposal, and then a deep network is used for feature extraction for each area. Then send the features to each SVM classifier for judgment, select a category of objects, find the candidate boxes with the highest score, calculate the IoU between it and similar region proposals, delete the region proposals with IoU greater than the threshold, and then filter Next category. Finally, since the region proposals generated using the SS method are not accurate, a regressor needs to be used to adjust the remaining region proposals. Of course, the first generation of R-CNN has many problems, such as the need to convolve all region proposals during feature extraction, redundant operations, or due to the existence of SVM, features need to be written to disk during training, resulting in training speed Slow, cumbersome process, etc.

So Fast R-CNN was born [9]. Fast R-CNN was also designed and created by Ross Girshick. On the basis of R-CNN [8], Fast R-CNN introduces the Region of Interest (ROI) pooling layer, so that no matter what the size of the feature matrix is, it can be uniformly scaled to a unified size, getting rid of the input image size. Limit. Compared with R-CNN, its training time is 9 times [9] faster, and test inference time is 213 times faster. On the Pascal VOC dataset [2], mAP increased from 62% to 66%, which is a great improvement. Despite this, the SS algorithm is still the bottleneck limiting the speed of Fast R-CNN.

In Faster R-CNN [10], SS algorithm is also handed over to the neural network for processing, and the Region Proposal Network (RPN) is implemented. When Faster R-CNN also uses VGG16 as the

backbone, the inference speed reaches 5fps on the GPU, and mAP has also been improved to a certain extent.

You Only Look Once (YOLO). Different from the two-stage detection of the R-CNN [8] series mentioned earlier, You Only Look Once (YOLO) [11] does not have the process of obtaining the region proposal. It treats the object detection process as a regression problem and only requires Look Once. In YOLOv1, the concept of single-stage detection was first proposed, taking the entire image as input and returning the position and category of the bounding box. This approach will make YOLO recognition very fast, a fast version of YOLO achieves a frame rate of 155 frames per second (fps) with a VOC07 mAP of 52.7%, which greatly outpaces the processing the speed of R-CNN [8] and Fast R-CNN [9] at the time. However, YOLOv1 also has shortcomings, such as reduced recognition accuracy for smaller or relatively close objects, or insufficient accuracy in judging the size of objects. In subsequent versions of YOLO, multi-scale prediction, residual blocks, skip connections, dynamic label assignment and model structure reparameterization, etc. were successively introduced, further improving the speed and accuracy of YOLO.

4. Conclusion

In terms of metrics, it mainly introduces indicators for measuring image recognition accuracy and speed, explains in detail the concepts of TP, FP, TN and FN, and derives a series of evaluation indicators, such as IoU, AP, mAP etc. In terms of data sets, two classic data sets, PASCAL VOC and DOTA, are introduced, and their applications in actual scenarios are described. In addition, this article also describes two important periods of object detection—the traditional object detection stage and the deep learning-based object detection stage. The iconic detectors in traditional target detection period include Viola-Jones Detector and HOG Detector. We introduced the R-CNN series and YOLO based on deep learning algorithms. In the article, we also review the performance and applicable scenarios of these detectors and algorithms.

Although the current object detection algorithms have achieved good results, there is still much room for progress in small object detection. Small object detection is an important issue in object detection field in the future. Since small objects have fewer pixels, RGB, and lower contrast, they have always been a difficult part of object detection. Multi-scale detection strategies, feature enhancement techniques, attention mechanisms and other methods can be used to solve the problem of small target detection. In practical problems, multi-object detection is very common. Since most object detection algorithms use the sliding window model, when there are multiple objects, the computational complexity will be greatly increased, and there may be problems such as object occlusion and object overlap. Multi-object detection problems can be solved by utilizing technologies such as associated information between objects, integrated deep learning, and traditional methods.

References

[1] Padilla R, Netto S L, Da Silva E A B. A survey on performance metrics for object-detection algorithms, 2020 international conference on systems, signals and image processing (IWSSIP). IEEE, 2020: 237-242.

[2] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, Andrew Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. International Journal of Computer Vision, 2010, 88: 303–338.

[3] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, Liangpei Zhang. DOTA: ALarge-scale Dataset for Object Detection in Aerial Images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 3974-3983.

[4] ZhengxiaZou, KeyanChen, ZhenweiShi, YuhongGuo, and JiepingYe. Object Detection in 20 Years: A Survey. Proceedings of the IEEE, 2023, 111(3): 257-276.

[5] Wang X, Han T X, Yan S. An HOG-LBP human detector with partial occlusion handling, 2009 IEEE 12th international conference on computer vision. 2009: 32-39.

[6] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation, Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[7] Dalal N, Triggs B. Histograms of oriented gradients for human detection. 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). 2005, 1: 886-893.

[8] Ross Girshick. Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015: 1440-1448.

[9] S. Ren, K. He, R. Girshick, and J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, 2015: 91–99.

[10] Kaiming He, Georgia Gkioxari, Piotr Doll´ar, Ross Girshick. Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017: 2961-2969.

[11] Yi-Qing Wang. An Analysis of the Viola-Jones Face Detection Algorithm. Image Processing On Line, 2014, 4: 128-148.