A Survey of Recent Trends in Two-Stage Object Detection Methods

M. F. Ansari and K. A. Lodi

Abstract Object detection deals with locating an object in an image or a video and identifying its class label. In this regard, integration of object detection techniques with deep learning has revolutionized the area of computer vision. Furthermore, the ability of a deep neural network to directly learn feature representation from images has significantly improved object detection models. The present survey analyzes and systematically provides a comprehensive overview of typical two-stage object detection methods with deep learning and summarizes the most popular benchmark dataset for object detection.

Keywords Object detection · Deep learning · Two-stage object detection

1 Introduction

Deep learning (DL) has empowered computer vision to effectively learn image features. It should be noted that most object detectors use a deep neural network as their backbone architecture to extract a feature from images and a detection network to detect objects in images or videos. An object detection method locates objects of a particular category in images or videos. It has fascinated researchers over the last decade. This technology has been applied to humans and society in the form of self-driving cars, face detection, activity recognition, pedestrian detection medical imaging, robotics, object counting, and crop monitoring. Recent developments in computational power have significantly contributed to the development of object detection techniques [1–8].

M. F. Ansari (✉)
Department of Computer Science, Aligarh Muslim University, Aligarh, India
e-mail: mfansari2395@gmail.com

K. A. Lodi
Department of Electrical Engineering, Aligarh Muslim University, Aligarh, India

A. Iqbal et al. (eds.), *Renewable Power for Sustainable Growth*, Lecture Notes
in Electrical Engineering 723, https://doi.org/10.1007/978-981-33-4080-0_64

Several benchmark datasets for instance KITTI, Caltech, MS COCO, PASCAL VOC, and Open Image V5 have played an important role in improving object detection. Organizations maintain a public dataset containing images and videos and information needed how to use them; anyone can download these datasets and conduct experiments. Presently, detection based on deep neural network can be divided into two classes:

- Two-stage detector
- One-stage detector.

R-CNN [1] and its different types are examples of two-stage detectors, whereas YOLO [2], its variants are one-stage detectors. Two-stage detectors are highly accurate in terms of localization and classification, whereas one-stage detectors have greater speed in terms of real-time detection. The stages of a two-stage detector can be specified; for instance, in faster R-CNN [4], first stage is called the region proposal network (RPN), which proposes a bounding box; in the next stage, features are pulled out from these boxes with the aid of the RoI pool (RoI pooling) operation. Architecture of a two-stage detector can be viewed in Fig. 1. One-stage detectors, on the other hand, directly predict bounding boxes from input images and the corresponding class label of each box. Architecture of a two-stage detector can be viewed in Fig. 2.

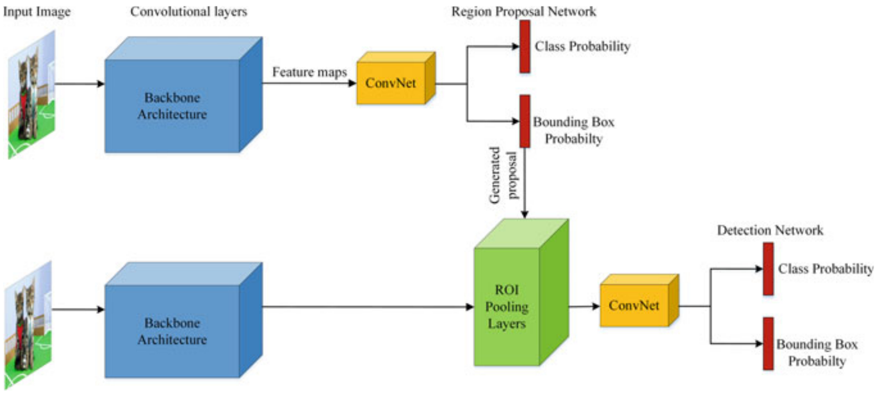

Fig. 1 Two-stage detector

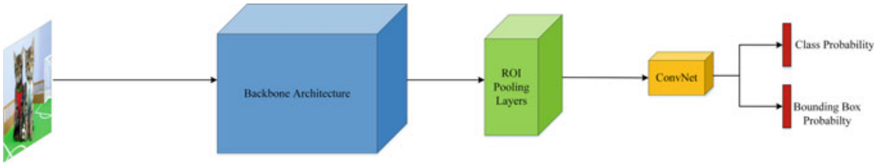

Fig. 2 One-stage detector

The main objective of this survey paper is to provide an all-inclusive understanding of DL-based two-stage object detection. The authors have reviewed numerous papers and their contribution to object detection; although many survey papers have been published on object detection on border way, but the literature lacks paper focusing on recent development and the recent start of art which have achieved great success in two-stage object detection. Furthermore, the authors have provided a summary on convolutional neural network (CNN) architecture, which serves as the backbone network for a feature extractor in the detection task and is described as the most popular two-stage detector. The authors have also summarized an explanation of popular benchmark datasets for object detection and evaluation metrics.

Section 2 presents the problem definition. Section 3 discusses the popular backbone architecture for object detection. In Sect. 4, authors have covered detail description of two-stage detection methods. Section 5 summarizes information about the application of object detection. Finally, conclusion is given in Sect. 6.

2 Problem Definition

Handcrafted features were one of the main limitations in terms of obtaining good accuracy computer vision tasks. However, with the rise of DL methods, the accuracy of solving vision problems has improved significantly. One of the major problems was object classification, which refers to categorizing all objects present in images into their respective classes. Object detection, also described as object category detection, is a more complex task than classification, as it involves predicting the class of a particular object and its precise location from an input image.

3 Popular Backbone Architecture

The primary requirement of good object detection is to learn a good feature representation. If the learned features are good enough, high accuracy in terms of object detection can be achieved. The popular backbone DCNN architecture widely used in object detection is AlexNet, VGGNet, ResNet, InceptionNet, and ResNeXt.

AlexNet was the first network architecture to be proposed by Krizhevesky in 2012 [3]. It possessed the ability to learn good representation from input images, with a minimal number (8) of layers. It has improved accuracy by a huge margin in the ILSVRC classification challenge [4]. VGG-16, with 16 layers, was based on AlexNet. After further increasing the number of the layers to 20 o network witnessed a dip in accuracy. In [5], the concept of a skip connection was introduced and the new ResNet was proposed, which reduced difficulties pertaining to optimization. This network can be extended to 100 layers with only a few parameters, as compared to VGGNet and AlexNet. Later, its various variants were proposed.

4 Detection Scheme Build on Deep Learning

In a two-stage detector, the first stage is used to generate the proposal in which potential objects can be present. During the second stage, predictions are made based on the generated proposal. The current two-stage detector can more accurately predict an object's location based on benchmark datasets.

4.1 R-CNN

R-CNN was the first network to be formed on CNN. After the success of the CNN in classification tasks, Ross Girshick proposed the R-CNN network for object detection. The R-CNN detects objects in three phases:

(i) Region generation phase
(ii) Extraction of feature phase
(iii) Prediction phase for classification and regression.

In the first phase, the R-CNN makes use of selective search algorithm (SRA) to select important regions in every input image; the selected regions are known as proposed regions. The advantage of using selective search is that it searches 2000 regions where objects can be present. In the second stage, the selected regions are cropped, resized, and fed into the CNN. At this phase, the CNN produces a 4096 dimensional feature vector as output. In the final step, classification and bounding box prediction happen. Architecture of the R-CNN can be viewed in Fig. 3.

The R-CNN considerably improved the object detection performance of traditional algorithms by a huge margin. However, it still has a few flaws:

(i) The extraction of features from the 2000 selected regions through a deep CNN requires a long computational time.
(ii) Optimization is difficult, as the network is divided into three stages.
(iii) Computational time is a lot for test images.

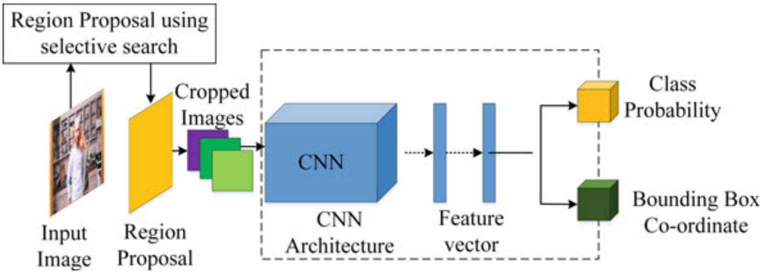

Fig. 3 R-CNN

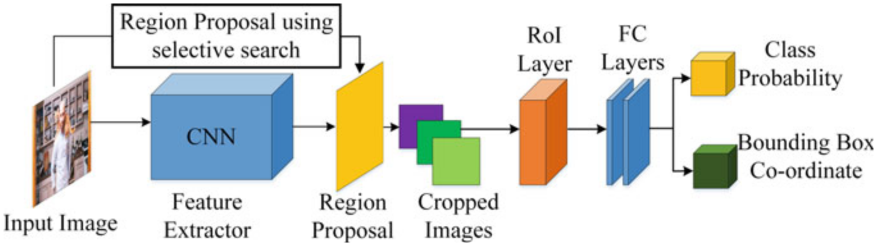

Fig. 4 Fast R-CNN

4.2 *Fast R-CNN*

After the addressing constraint of the SPPNet and R-CNN, Ross Girshick et al. proposed a fast detection algorithm called fast R-CNN. It is same as the R-CNN, except that the generated region is fed to the CNN. It takes the whole image as input and feeds it to the CNN to obtain the convolutional features map. After convolution, the feature map goes from the RoI layer, which generates the reshaped feature with a fixed size. Fixed features are fed to the classification and regression layers to predict class labels and bounding boxes, respectively. Fast R-CNN extracts features from entire images, whereas R-CNN uses 2000 regions to extract features. This saves immense time during training and testing. Architecture of the fast R-CNN can be viewed in Fig. 4.

4.3 *Faster R-CNN*

Both previous networks were based on traditional SRA, which were slow, time-consuming, and capture only low-level features in the features map. Faster R-CNN, developed by [4], uses the RPN to generate regions based on the CNN. The RPN generates regions from input images by feeding them into the CNN. It also increases the generation of region proposals with the aid a common set of CNN layers with detector network. After being generated, the regions are changed using RoI pooling layer. The image is then fed to the classification and regression layer for label classification and offset prediction. Faster R-CNN achieved relatively better results with respect to object detection benchmark datasets for instance MSCOCO, Pascal VOC, and ILSVRC [5]. The stages of the network have been outlined in Fig. 5.

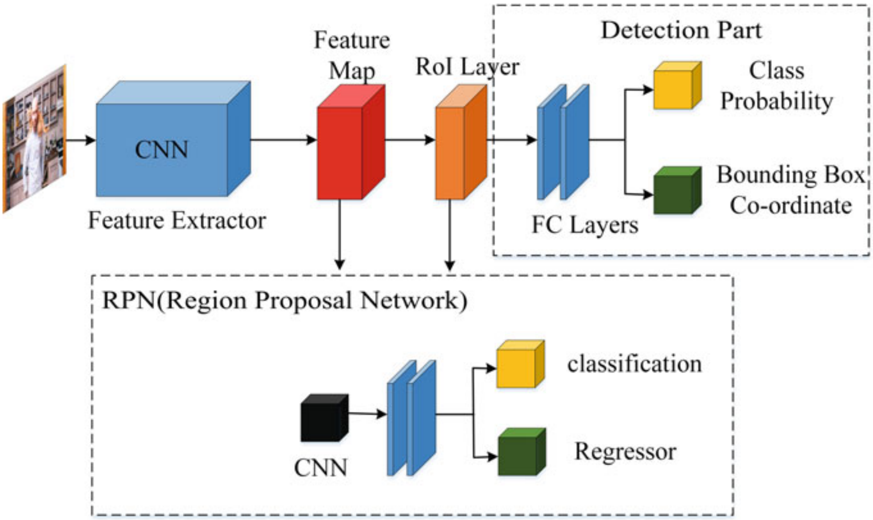

Fig. 5 Faster R-CNN

4.4 R-FCN

In R-FCN [5], the fully connected (FC) layers that follow the RoI layers are removed, and all major complex features are assigned before the RoI layers. The R-FCN generates position-sensitive maps, which contain information about position regarding distinct classes. The position-sensitive RoI layer is applied to pull out features from score maps. The R-FCN makes use of simple average voting on extracted features from the RoI layer to generate a class vector. At last, the Softmax function is performed on this vector to predict the class score. Architecture of the R-FCN can be viewed in Fig. 6.

4.5 Mask R-CNN

For pixel level detection, He et al. [6] developed the instance segmentation algorithm, the mask R-CNN. This can be viewed as an extension of the faster R-CNN. The Mask R-CNN uses a two-phase strategy. In the first phase, it uses the RPN to generate regions where objects might be present. In the second phase, it foresees the binary mask based on the feature map. A mask-generating branch based on CNN is used to better capture the relevant areas. The mask R-CNN uses RoI align layer, in place of the RoI layer with backbone architecture. Mask R-CNN is simple to accomplish and achieves better accuracy in terms of the instance segmentation task. Figure 7 shows the architecture of the mask R-CNN.

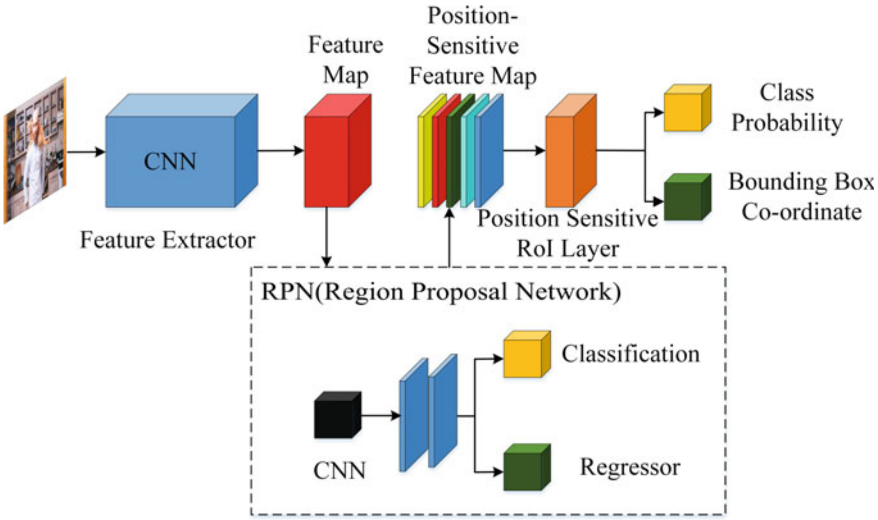

Fig. 6 R-FCN

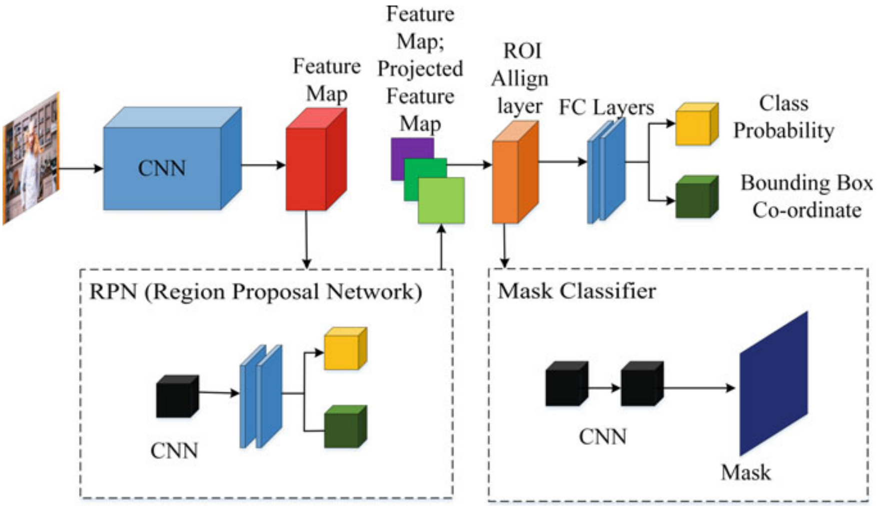

Fig. 7 Mask R-CNN

5 Object Detection Application

There are wide range object detection application in real-world scenarios, spanning from social to personal levels (Fig. 8).

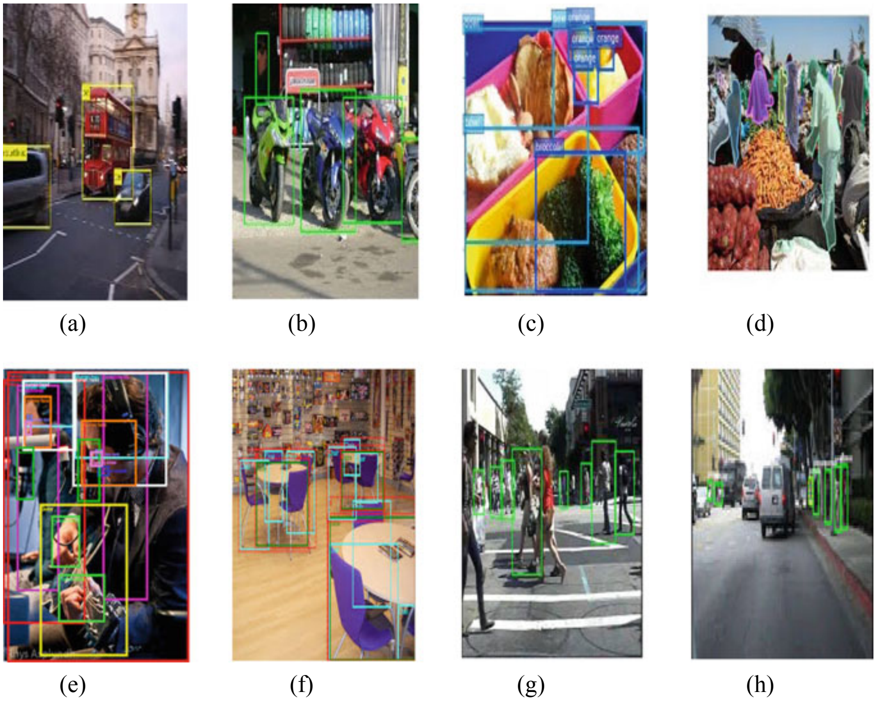

Fig. 8 Benchmark datasets, example **a**, **b** are from Pascal Voc dataset, example **c**, **d**, are from MS COCO dataset, **e**, **f** are from Open Image v5, example **g**, **h** from Caltech dataset

- **Face detection**: Detection of faces is a very prominent area in computer vision; it involves detecting human faces from images or videos. It has many applications for instance in security, health care, advertisements, and so on.
- **Pedestrian Detection**: It is worth noting that several specific datasets have been published on pedestrian detection. The Euro City Persons dataset, for example, contains information regarding pedestrians, cyclists, and other riders in traffic areas.
- **Text Detection**: Text detection deals with detecting text area in images or videos text detection have many applications for example in identifying vehicles by reading number plates, in assisting visually impaired persons.

6 Conclusion

Over the last few years, with the advancement of DL, object detection tasks have evolved rapidly. In this survey, the authors reviewed the modern literature on object detection, covering all relevant information about two-stage object detection and describing backbone architecture. The authors also covered the popular benchmarks of object detection and evaluation matrix. The authors even attempted to cover all terminologies in a deterministic manner to allow the survey to better compress object detection based on deep learning.

References

1. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of IEEE computer society conference on computer vision and pattern recognition, pp 580–587. https://doi.org/10.1109/CVPR.2014.81
2. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: Proceedings of IEEE computer society conference on computer vision pattern recognition, pp 779–788. https://doi.org/10.1109/CVPR.2016.91
3. Girshick R (2015) Fast R-CNN. In: Proceedings of IEEE international conference on computer vision, ICCV 2015, pp 1440–1448. https://doi.org/10.1109/ICCV.2015.169
4. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: NeurIPS
5. Dai J, Li Y, He K, Sun J (2016) R-FCN: Object detection via region-based fully convolutional networks. In: Advances in neural information processing systems, pp. 379–387
6. He K, Gkioxari G, Dollár P, Girshick R (2020) Mask R-CNN. IEEE Trans Pattern Anal Mach Intell 42:386–397. https://doi.org/10.1109/TPAMI.2018.2844175
7. Iqbal A et al (eds) (2020) Soft computing in condition monitoring and diagnostics of electrical and mechanical systems. In: Advances in intelligent systems and computing, vol 1096. Springer, Singapore. https://doi.org/10.1007/978-981-15-1532-3
8. Iqbal A et al (eds) (2020) Meta heuristic and evolutionary computation: algorithms and applications. In: Studies in computational intelligence, vol 1096. Springer, Singapore. https://www.springer.com/gp/book/9789811575709