Original Research Article

A Survey of object detection in crowded scenes based on deep learning

Haonan Tian

Hunan University of Science and Technology, School of Computer Science and Engineering, Xiangtan, 411100, China

Abstract: With the progress of deep learning technology, the object detection algorithms have achieved good detection results in general scenes, but they have encountered difficulties in crowded scenes. In crowded scenes, there are a lot of occlusion between objects, which makes the non-maximum suppression algorithm easy to delete the correct detection of overlapping objects; at the same time, there are some problems, such as large change of object scale, small object, insufficient available features and so on. In order to promote the further development of crowded object detection technology, the related methods and techniques are summarized. Firstly, the research background and application value of object detection in crowded scenes are introduced. Secondly, the difficulties of object detection in crowded scenes are discussed, and the defects of object detection algorithm based on deep learning in crowded scenes are analyzed. Then, the existing object detection algorithms in crowded scenes are classified and described. Finally, the possible research direction of object detection in crowded scenes is prospected.

Keywords: Deep learning; Object detection; Crowded scenes; Non-maximum suppression; YOLO

1. Introduction

Object detection is a fundamental task in the field of computer vision, aimed at identifying all objects of interest within an image. Unlike simple image classification tasks, object detection requires identifying multiple objects within an image, determining their categories, and providing specific location information using rectangular bounding boxes. With advancements in computer hardware performance and the development of technologies such as deep learning, object detection has made significant progress in various fields and is widely used in applications including security surveillance, medical image analysis, autonomous driving, and industrial inspection[1]. The powerful learning capabilities of deep neural networks have led to substantial success in the accuracy and efficiency of general object detection. However, in certain scenes, it still falls short of human recognition abilities. In crowded scenes, such as train stations, airports, and shopping malls, object detection algorithms often exhibit a noticeable decline in performance. Detecting objects in crowded scenes can help in monitoring traffic flow or preventing potential security threats, which demonstrates significant practical application value.

In crowded scenes, the accuracy of general object detection algorithms significantly decreases and fails to meet the expected or required performance. To address this issue, many researchers have developed numerous object detection algorithms specifically for crowded scenes based on deep learning, which have achieved fruitful improvements. This paper reviews the difficulties in the task of object detection in crowded scenes, and, by referencing recent research results, summarizes the main improvement methods.

2. Difficulties in crowded scenes

Object detection technology based on deep learning has made significant strides; however, detection tasks in crowded scenes remain a formidable challenge. As illustrated in Figure 1, the primary difficulties of object

detection in crowded scenes include the following key aspects:

（a）occluded objects （b）small objects （c）insufficient features

Figure 1. Examples of different difficult scenes.

(1) The objects are severely occluded. Due to the presence of numerous objects in crowded scenes within a single image, they are positioned very close to each other, which results in significant occlusion among them. The mutual occlusion of objects of the same category can lead to situations where, even if the detection algorithm correctly generates detection boxes for the occluded objects, these detection boxes may be removed by the post-processing algorithm as duplicate detections, thereby causing missed detections.

(2) There are numerous small objects. In some outdoor open scenarios, the objects are generally at a considerable distance, which results in small scales in the images. Small objects have low resolution, contain few pixels, and have very blurry shapes. This makes them difficult to be recognized. During the downsampling process in the network, there is also a loss of detailed information, which leads to even fewer features of small objects in higher-level feature maps. This further hinders the detection of small objects.

(3) The object features are insufficient. Crowded object detection encounters highly complex scenarios, where objects or adverse lighting conditions may negatively impact object imaging. Both intra-class occlusion and occlusion by non-target objects in the scenes can lead to insufficient visible parts of the object, thereby affecting object detection. Intense light, low light, and rapid object movement can cause object blurring and sufficient features, which hampers detection.

3. Non-maximum suppression

An end-to-end object detection algorithm should generate a unique detection result for each real object. Therefore, it needs to handle multiple detection boxes for the same object, retain the detection box with the highest confidence score, and eliminate other redundant detection boxes. This processing method is known as Non-Maximum Suppression (NMS). Greedy-NMS is the standard algorithm used for NMS in object detection algorithms. It is simple and effective for general object detection tasks and is the most commonly used post-processing method.

The effectiveness of Greedy-NMS is based on the assumption that other detection boxes with high overlap around the detection box with the highest confidence score are redundant predictions of the same real object. However, this is not accurate in crowded scenes. In crowded scenes, real objects often occlude each other, which results in highly overlapping detection boxes for nearby objects. This can cause some correctly detected boxes to be erroneously suppressed by NMS. As illustrated in **Figure 2**, because object a and object b are very close to each other, the detection boxes generated for them by the object detection algorithm also have a high degree of overlap. During the NMS process, when a certain detection box for the object a is selected as the highest scoring box, all detection boxes for object b are deleted as redundant detections due to their Intersection over Unions (IoUs) with this detection box for object a being higher than the suppression threshold. This leads to the missed detection for object b.

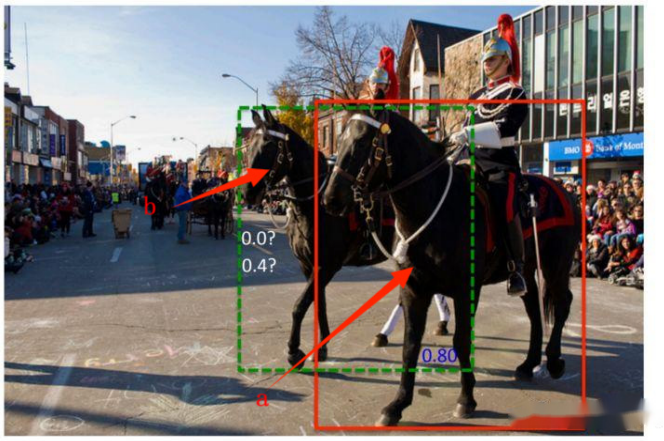

Figure 2. The schematic diagram of erroneous suppression.

4. Crowded object detection algorithms

Object detection in crowded scenes is a challenging task in the field of computer vision, and its research aims to achieve accurate classification and localization of objects in highly complex and crowded scenes. Scholars have made a series of improvements based on general object detection algorithms to address object detection in crowded scenes. These improved algorithms can be roughly divided into two categories: crowded object detection algorithms based on improving NMS effects and crowded object detection algorithms based on improving the models.

4.1. Crowded object detection algorithms based on improving NMS effects

Greedy-NMS is a standard post-processing method for object detection algorithms and performs well in general object detection tasks. However, it often produces erroneous suppression when dealing with crowded objects. Some efforts have been made to improve the NMS algorithm itself. Soft-NMS[2] proposed by Bodla et al. does not directly remove overlapping detection boxes but adopts a score decay strategy. It reduces the detection scores of these overlapping boxes gradually based on their IOUs with the highest scoring box, thus gradually eliminating redundant detection boxes. However, this strategy retains more detection boxes in each round of suppression, severely impacting the efficiency of the NMS algorithm. Adaptive-NMS[3] provides a novel NMS scheme for crowded scenes. It sets an additional density sub-network at the higher layers of the detection network. This sub-network learns the object density in different regions of the image and provides adaptive IOU thresholds for NMS, which can alleviate the difficulty of setting IOU thresholds in crowded scenes.

Some improved algorithms not directly modifies NMS itself but designs special loss functions for crowded scenes to generate more compact detection boxes, thereby improving the performance of standard NMS. Repulsion loss[4] is proposed to address the challenge of occlusion in crowded scenes, which consists of one attraction term and two repulsion terms. The attraction term is used to make the predicted boxes closer to their designated objects, while the two repulsion terms are used to push the predicted boxes away from other surrounding true objects and predicted boxes from other objects. Aggregation loss[5] designed by Zhang et al. aims to make the predicted boxes from the same true object more compact, thereby reducing erroneous suppression by NMS when dealing with occluded objects.

4.2. Crowded object detection algorithms based on improving the models

Some object detection algorithms for crowded scenes have been improved in terms of model structure or

feature extraction modules, which aims to generate high-quality prediction results to enhance detection capabilities in crowded scenes. Chu et al. propose a multi-instance prediction method to address the issue of overlapping objects in crowded object detection. This method generates a set of highly overlapping predictions for each anchor box, thereby increasing the likelihood of predicting crowded objects. Rukhovich et al. introduces a multi-round iterative detection algorithm. It first detects simple objects in the image, then uses the detection results as feature maps and re-inputs them into the neural network along with the low-level feature maps of the original image. This allows the network to further explore difficult objects that have not been detected based on historical detection results. Zheng et al. employe a query-based approach for object detection in crowded scenes and propose a progressive detection algorithm. This algorithm first treats high-confidence query results as acceptable queries, and then uses these accepted queries to determine whether the remaining noisy queries have detected real objects.

Some researchers explore real-time crowded object detection algorithms by improving the YOLO series detection algorithms. Gao et al. enhance the network structure based on YOLOv5, introduce new modules and propose the V-YOLO detection algorithm[6]. This algorithm uses a bidirectional weighted feature pyramid network instead of the original path aggregation network to achieve better feature fusion effects by reasonably setting the paths for feature propagation. Liu et al. propose a lighter-weight Bi-YOLO detection algorithm[7], which is based on YOLOv8 and introduces the GSConv module to further reduce parameters. Bi-YOLO introduces a dynamic sparse attention module, BiFormer, based on double-layer routing. This module can improve issues such as computational complexity and large memory footprint of the Transformer structure. It adaptively focuses on the content of relevant regions through query perception.

5. Conclusion

In the research on crowded object detection, researchers primarily focus on improving general object detection models to address issues such as error suppression caused by occlusion and missed detections. They achieve certain effectiveness. However, challenges and unresolved issues still persist in the research. In the future, object detection in crowded scenes may evolve in the following directions:

(1) Fine-grained feature representation and learning: By introducing richer semantic information, spatial information, and contextual information, researchers design more refined feature representation methods to enhance the performance of object detection in crowded scenes.

(2) Modeling relationships between objects: Researchers considers the interrelationships between objects, such as occlusion relationships, aggregation relationships, etc. and utilizes methods like graph neural networks to model the relationships between objects, thereby improving the accuracy of object detection.

(3) Lightweight and real-time: Researchers designs efficient model structures and optimization algorithms to reduce the computational complexity of object detection in crowded scenes and achieve real-time detection on embedded or mobile devices.

(4) Self-supervised learning and weakly supervised learning: By utilizing methods of self-supervised learning and weakly supervised learning to learn object detection models from large-scale unlabeled or weakly labeled data, researchers can mitigate the problem of insufficient annotated data.

The research on object detection in crowded scenes faces numerous challenges. Through continuous exploration of new methods and technologies, it is hoped to further improve the accuracy and robustness of crowded object detection algorithms and better meet the demands of object detection in complex real-world scenarios.

About the author

Haonan Tian is a master's student at the School of Computer Science and Engineering, Hunan University of Science and Technology. His research focuses on object detection in crowded scenes.

References

[1] Zou Z, Chen K, Shi Z, et al. Object detection in 20 years: A survey[J]. Proceedings of the IEEE, 2023, 111(3): 257-276.

[2] Bodla N, Singh B, Chellappa R, et al. Soft-NMS--improving object detection with one line of code[C]// Proceedings of the IEEE international conference on computer vision. 2017: 5561-5569.

[3] Liu S, Huang D, Wang Y. Adaptive nms: Refining pedestrian detection in a crowd[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 6459-6468.

[4] Wang X, Xiao T, Jiang Y, et al. Repulsion loss: Detecting pedestrians in a crowd[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7774-7783.

[5] Zhang S, Wen L, Bian X, et al. Occlusion-aware R-CNN: detecting pedestrians in a crowd[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 637-653.

[6] Gao Q, Tang F, Li D, et al. Research on pedestrian detection method in dense scene based on improved YOLOv5. Foreign Electronic Measurement Technology. 2023;42(125-130).

[7] Liu Z, Xu H, Zhu X, et al. Bi-YOLO: An Improved Lightweight Object Detection Algorithm Based on YOLOv8. Computer Engineering Science.(1-15).