Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/20/20231066

Analysis based on object detection algorithms

Puran Lyu

College of Arts and Science, New York University, 32 Waverly Pl, New York, NY 10003, USA

pl2355@nyu.edu

Abstract. With the increasing demand for intelligent systems capable of comprehending visual information, the discipline of image object detection has experienced rapid expansion. Despite the fact that numerous methods have been proposed, the existing literature lacks exhaustive analyses and summaries of these methods. This paper seeks to address this deficiency by providing a thorough overview and analysis of image object detection techniques. This paper analyzes and discusses traditional methods and deep learning-based methods, with a focus on analyzing the current state and shortcomings of traditional methods. Further discussion is given to deep network-based object detection methods, mainly through a comparative analysis of two-stage and one-stage methods. The basic performance of the You Look Only Once (YOLO) series methods is highlighted. The contribution of large-scale datasets and evaluation metrics to the advancement of the state of the art is also examined. This comprehensive analysis is a useful reference for researchers who aim to contribute to the continual progress of image object detection.

Keywords: object detection, deep learning, YOLO, evaluation metrics.

1. Introduction

The demand for intelligent systems capable of comprehending and interpreting visual information has fuelled the explosive growth of computer vision over the past several decades. The problem of image object detection entails accurately identifying and locating objects within images or sequences of images. Object detection is the basis for numerous applications, including autonomous vehicles, robotics, video surveillance, medical image analysis, and augmented reality [1].

Historically, object detection methods relied on the hand-crafted feature, limiting their robustness, scalability, and adaptability to new object classes or diverse environments. Viola-Jones algorithm for face detection, Scale-Invariant Feature Transform (SIFT) for key point detection and matching, and Histogram of Oriented Gradients (HOG) for person detection are examples of classical object detection techniques. While these techniques were instrumental in the early development of object detection, they have been vastly eclipsed by more recent techniques based on deep learning. The emergence of deep learning and the increasing availability of large-scale annotated datasets have resulted in a significant paradigm shift in object detection. Convolutional neural networks (CNNs) have been extensively adopted to automatically learn hierarchical feature representations directly from unprocessed pixel values, eliminating the need for manual feature engineering. This has led to the creation of a number of cutting-edge object detection techniques [2].

DOI: 10.54254/2755-2721/20/20231066

Despite the significant advancements made in recent years, numerous obstacles remain in the field of image object detection. Researchers continue to contend with issues such as class imbalance, detecting small objects, addressing occlusion and clutter, and ensuring domain shift robustness. In addition, emerging trends in object detection, such as anchor-free methods, transformer-based architectures, and few-shot detection techniques, hold promise for addressing some of these obstacles and expanding the field's capabilities.

Extensive approaches to these problems have been proposed, but comparative analyses and summaries of these approaches are lacking in the literature. This article provides a comprehensive overview of image object detection techniques, including classical approaches, methods based on deep learning, and emerging trends. The role of large-scale datasets, evaluation metrics, and benchmarking platforms in advancing the state-of-the-art is investigated. By providing a comprehensive comprehension of the current state of image object detection, its challenges, and future directions, we hope to provide a valuable resource for researchers who wish to contribute to the continued advancement of this rapidly advancing field.

2. Classical approaches to object detection

In the early days of computer vision, object detection relied on classical approaches involving hand-crafted feature extraction, such as HOG and SIFT, followed by the use of machine learning classifiers such as Support Vector Machine (SVM) or Bayesian classifiers for detection. The HOG descriptor is a feature extraction method that captures the distribution of gradient orientations in localized regions of an image [3] (Figure 1). It divides the image into small cells, computes histograms of gradient orientations within each cell, and then normalizes these histograms over larger blocks to account for variations in lighting and contrast. The final HOG feature vector is composed of concatenated normalized histograms and can be used to train a classifier such as a SVM for object detection. HOG has proven especially useful for human detection, as it can capture the shape and structure of human silhouettes.

However, there are limitations to the HOG descriptor. It might not be resistant to changes in object appearance, viewpoint, or occlusions. Moreover, the process of hand-crafting features can be time-consuming and may require significant domain knowledge and experimentation [4].

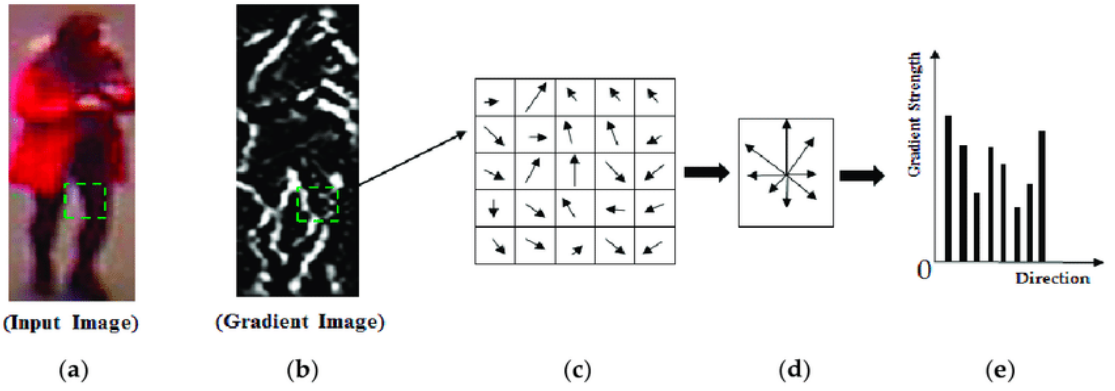

Figure 1. HOG feature extraction process.

In addition to HOG, another classical method that has made a significant impact on object detection is SIFT. SIFT is a prominent feature extraction method that identifies and describes the scale and rotation-invariant key points within an image. The SIFT algorithm requires the detection of scale-space extrema, the assignment of orientations to key points, and the generation of a descriptor for each key point based on local image gradients. SIFT descriptors can be matched across multiple images to identify corresponding object instances or used to train classifiers for object detection. SIFT has limitations when it comes to object appearance, lighting conditions, and affine transformations, despite its resistance to changes in scale and rotation. Moreover, its computational complexity can be problematic for real-time

DOI: 10.54254/2755-2721/20/20231066

applications. There are methods for object detection that combine SIFT and classifiers to enhance detection performance. For instance, when SIFT and SVM are combined, SIFT features are extracted from the input images and then used to train a SVM classifier. The classifier is then applied to new images to detect objects. This combination of SIFT and SVM has been applied to a variety of object detection tasks with promising outcomes [5].

While SIFT has been a popular choice for object detection, several enhancements and variants have been proposed to boost its efficacy. The incorporation of spatial verification techniques to refine object localization is one such enhancement. After initial object detection with SIFT key points, additional geometric constraints or spatial consistency tests can be used to validate and refine the object's bounding box. This serves to reduce false detections and improve object localization precision. Integration of deep learning techniques with SIFT constitutes an additional enhancement (Figure 2). Deep learning models, such as CNNs, have demonstrated remarkable capabilities for directly learning complex and discriminative features from raw pixel data. Combining the descriptive power of CNNs with the robustness of SIFT enables more precise and robust object detection. This hybrid strategy exploits the complementary advantages of both methods to improve detection performance. Overall, SIFT-based object detection methods, whether employed in their traditional form or augmented with additional techniques, have proven effective in a variety of object detection applications. They provide a solid foundation for feature extraction and classification, paving the way for future advances in object detection.

However, the handcrafted nature of these features and their reliance on engineered representations restricted their robustness, scalability, and adaptability to new object classes or different environments.

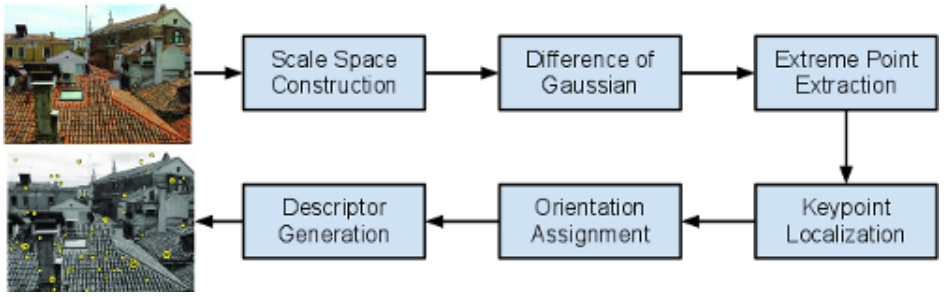

Figure 2. Flow chart of sift feature extraction.

3. Deep learning-based techniques

Deep learning has revolutionized the field of computer vision by facilitating the extraction of features automatically from unprocessed pixel values. This has significantly reduced the need for hand-crafted features and engineered representations, resulting in substantial advancements in object detection performance. CNNs are the most extensively used deep learning architectures for image object detection, as they offer both end-to-end training and hierarchical feature learning [6]. Current approaches can be categorized broadly into one-stage and two-stage object detections, which will be elaborated upon in the following sections.

3.1. Two-stage object detection

Two-stage object detection techniques are distinguished by the use of region proposition techniques followed by CNN feature extraction and classification. Typically, these methods involve three primary steps: region proposal generation, feature extraction using CNNs, and object classification using classifiers such as SVM.

The first stage of two-stage object detection entails the generation of a set of region proposals, which are possible image object bounding boxes. Different strategies, such as selective search, edge boxes, and region proposal networks (RPNs), have been used to generate region proposals. The objective is to generate a set of diverse candidate regions that are likely to contain objects of interest. In the second phase, a CNN that has been previously trained is used to derive features from the proposed regions. The

DOI: 10.54254/2755-2721/20/20231066

CNN is typically trained on a large dataset for image classification before being fine-tuned for the specific object detection task. The extracted features capture hierarchical and spatial information about the objects, providing a comprehensive and discriminative representation for classification. Finally, a classifier such as an SVM is applied to the extracted features to determine the object class and refine the bounding box coordinates. The classifier is trained using ground-truth annotations of object classes and their corresponding bounding boxes in the training images. Some two-stage object detection methods replace the SVM classifier with a fully connected layer within the CNN architecture, allowing for end-to-end training of the entire model.

Region-based CNN (RCNN)is an example of a two-stage object detection method. R-CNN employs a selective search for region proposal generation, extracts CNN features from the proposed regions, and then uses an SVM for object classification (Figure 3). This method has demonstrated a substantial improvement in object detection performance compared to conventional methods [7]. Fast R-CNN is an enhanced variant of the R-CNN algorithm that addresses some of its shortcomings, namely its sluggish performance. Fast R-CNN introduces a number of significant enhancements to increase performance and effectiveness. Instead of extracting features independently for each proposed region, Fast R-CNN extracts features from the entire image only once. This is accomplished through the use of a region of interest (RoI) pooling layer, which extracts fixed-size feature maps for each proposed region. Then, these feature maps are fed into layers with complete connections for classification and bounding box regression. Fast R-CNN considerably accelerates the process in comparison to the original R-CNN by sharing convolutional features across all proposals. Mask R-CNN is another example. It expands the capabilities of object detection to include instance segmentation, which requires not only the detection of objects but also the accurate delineation of their boundaries. Mask R-CNN generates a mask at the pixel level for each detected object in addition to predicting object classes and bounding box coordinates. This allows for the precise segmentation of image objects. Mask R-CNN accomplishes this by augmenting the Fast R-CNN architecture with a parallel mask branch. The mask branch is responsible for generating a binary mask that specifies the precise shape and location of each instance of an object.

By integrating the advantages of object detection and instance segmentation, Mask R-CNN provides more detailed and fine-grained object information in an image. This is beneficial in situations where objects may overlap or obscure one another. Image segmentation, instance-aware image modification, and real-time video object segmentation are examples of applications where Mask R-CNN has proven to be highly effective.

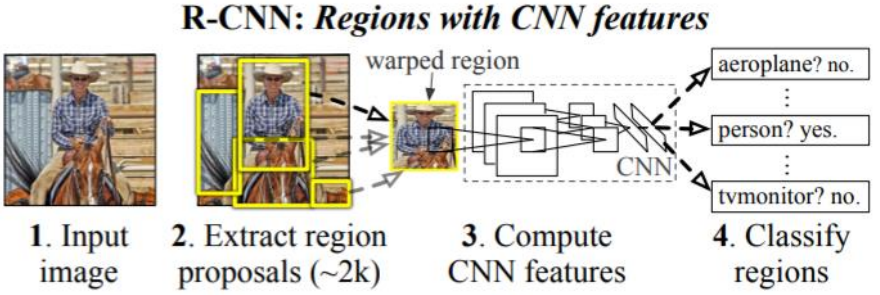

Figure 3. R-CNN flowchart.

However, there are disadvantages to two-stage object detection methods. Due to the separate phases of region proposal and feature extraction, they can be computationally expensive, limiting their applicability in real-time settings. Furthermore, reliance on region proposal techniques can introduce additional hyperparameters and complexity to the overall detection pipeline.

Two-stage object detection methods attain state-of-the-art performance in object detection by combining region proposal generation, CNN feature extraction, and classification. Although these approaches have considerably advanced the field, their computational complexity and reliance on region

DOI: 10.54254/2755-2721/20/20231066

proposal techniques have prompted the development of more efficient alternatives, such as one-stage object detection.

3.2. One-stage object detection

The detection process is accelerated in one-stage object detection methods as they predict object classes and bounding box coordinates in a single network iteration. This substantially improves computational efficiency and enables real-time object detection by eliminating the need for separate stages of region proposal generation and classification. Approaching object detection as a regression problem, YOLO is a revolutionary one-step approach that divides the input image and predicts the object categories and bounding box coordinates for each grid cell. By employing a single network to execute the entire detection task, YOLO significantly reduces computation time in comparison to two-stage methods.

Several enhanced variants of YOLO have been proposed since its inception. Various modifications and optimizations have been implemented in these versions to improve detection performance and efficiency. The use of anchor boxes, additional convolutional layers, batch normalization, and novel loss functions are among the most significant enhancements. In general, each new iteration of YOLO offers improved performance and faster inference times than its predecessor, making them more suitable for real-time applications. Among the main significant contributions and enhancements of various versions of YOLO are:

YOLOv1: The original version of YOLO introduced the concept of object detection as a regression problem using a singular neural network. It utilized a straightforward architecture consisting of 24 convolutional layers and 2 fully connected layers. In addition, a custom loss function was used to balance localization and classification errors.

YOLOv2: An enhanced version of YOLO that includes performance-enhancing techniques such as batch normalization, anchor boxes, multi-scale training, and fine-grained features. In addition, it utilized a new architecture with 23 convolutional layers and a passthrough layer that merged features from prior layers. At the time, its performance on PASCAL VOC (Pattern Analysis Statistical Modelling and Computational Learning, Visual Object Classes) and COCO (Microsoft Common Objects in Context) datasets was unparalleled.

YOLOv4 is an updated version of YOLO that incorporates numerous components from recent studies on network design, training strategies, testing techniques, self-adversarial training, and mosaic data augmentation [7]. Additionally, a new architecture with weighted residual connections, cross-stage partial connections, spatial attention modules, and mish activation functions were implemented. It outperformed previous versions on a variety of performance metrics.

YOLOv5 is a refined version of YOLO that uses the PyTorch framework as opposed to the Darknet framework. Additionally, a simplified architecture with fewer layers and parameters was utilized. It introduced new models at various grades (S, M, L, X) that combined speed and precision. It evaluated its efficiency and effectiveness across multiple datasets utilizing distinct metrics.

The most recent member of the YOLO family. The YOLOv8 framework is based on the YOLOv5 framework and includes numerous architectural and developer experience enhancements. It is faster and more accurate than YOLOv5, and it provides a unified framework for training object detection, instance segmentation, and image classification models [8].

As the field of image object detection continues to evolve, there are a number of noteworthy emerging trends to consider. These include anchor-free detection methods, transformer architectures, and few-shot detection techniques. Anchor-free methods seek to simplify the detection pipeline by removing the requirement for anchor boxes, which can introduce additional hyperparameters and complexity. Driven by the achievements of transformers in natural language processing, transformer architectures employ self-attention mechanisms to model long-range dependencies and capture spatial information more effectively. Few-shot detection techniques seek to train models capable of detecting novel object classes using only a few annotated examples, thereby addressing the problem of limited annotated data for certain object categories.

DOI: 10.54254/2755-2721/20/20231066

4. Metrics for evaluation and benchmarking

4.1. Measurement criteria

Several evaluation metrics are frequently employed to assess the effectiveness of object detection methods. Intersection over Union (IoU) is a commonly used metric that evaluates the overlap between the predicted and actual bounding boxes. Higher IoU values indicate superior performance in localization. Additionally, precision, recall, average precision, and average recall are used to evaluate the classification performance of the detection methods. In addition, the mean Average Precision (mAP) metric integrates precision and recall values across various object classes to provide a single performance metric that facilitates method comparison. Besides, the inference speed (frames per second, FPS) is also a significant evaluation. The greater the FPS value, the more rapidly the model is detected and the more real-time it appears. In general, the performance of the model is determined by combining the mean Average Precision and the inference speed.

4.2. Comparative analysis

The table 1 compares the performance of various object detection methods based on the same dataset using different evaluation metrics.

Table 1. Compares the performance of various methods based on different evaluation metrics.

Category	Method	Speed /(frame·s-1)	Microsoft Common Objects in Context (MS COCO) (mAP@IoU=0.5:0.95)	The PASCAL VOC Challenge 2012 (VOC2012) (mAP@IoU=0.5)	mAP(%)
Two-Stage	R-CNN	0.03		53.3%	58.5
	Fast R-CNN	3	19.7%	68.4%	70.0
	Mask R-CNN	11			78.2
Yolo series	YOLO v1	45		57.9%	57.9
	YOLO v2	40	21.6%	73.4%	73.5
	YOLO v3	20	33.0%		57.9
	YOLO v4	33	43.5%	42.1%	43.5
	YOLO v5n	30.826	28.0%		
	YOLO v6n	26.55	35.9%		
	YOLO v7t	20.006		52.8%	

Deep learning-based techniques typically outperform conventional approaches. This is primarily due to the limitations of hand-crafted features, which frequently fail to depict the complex and diverse appearance of real-world objects. Moreover, these approaches are sensitive to alterations in object appearance, lighting conditions, and affine transformations. Although two-stage methods provide

DOI: 10.54254/2755-2721/20/20231066

significant performance enhancements over conventional techniques, their inference speeds are typically slower. Mask R-CNN obtains the highest performance among two-stage approaches at the expense of speed. Methods based on YOLO are designed to detect objects in real-time while preserving competitive performance. One-stage methods in the YOLO family allow for a faster processing pace, but in some cases may sacrifice precision.

In conclusion, the selection of an object detection method depends on the application and needs at hand. When real-time detection is crucial, YOLO-based methods are preferable due to their rapid processing speed. Nonetheless, if precision is the primary concern, a two-stage method such as Mask R-CNN is preferable. In situations where techniques based on deep learning are impracticable due to resource constraints or other constraints, classical methods may still be applicable.

5. Conclusions

This article delves into the analysis and discussion of both traditional methods and deep learning-based methods, with a particular focus on examining the current state and limitations of traditional methods. The article further explores deep network-based object detection methods, providing a comparative analysis of two-stage and one-stage methods. The article also emphasizes the basic performance of the YOLO series methods. Additionally, the article examines the contribution of large-scale datasets and evaluation metrics to the advancement of the state of the art. In conclusion, traditional object detection approaches, such as HOG and SIFT, relied on hand-crafted features, which limited their robustness, scalability, and adaptability. With the advent of deep learning, the paradigm for object detection has shifted significantly towards automatic feature extraction using CNNs. Two-stage object detection methods, characterized by region proposal generation, CNN feature extraction, and classification, have shown state-of-the-art performance on a variety of object detection tasks. However, their computational complexity and reliance on region proposal techniques have prompted the investigation of more effective alternatives, such as one-stage object detection techniques.

Emerging trends, such as anchor-free detection, transformer architectures, and few-shot detection, hold promise for further enhancing the performance of object detection and addressing the limitations of current methods. As research in this area continues, we can anticipate the development of more sophisticated and robust object detection techniques suited to a broader spectrum of applications, including real-time and resource-constrained scenarios.

References

[1] Rath, S., & Gupta, V. Yolov5 vs yolov6 vs yolov7: Comparison of Yolo models on speed and accuracy: CPU & GPU. 2022, LearnOpenCV.

[2] Boesch, G. Yolov7: The most powerful object detection algorithm. 2023, viso.ai. https://viso.ai/deep-learning/yolov7-guide/

[3] MLBoy. Train yolov 5 with your own data. 2022 Medium 4.

[4] Deng, M., & Zhang, D. Survey on Deep Neural Network Image Target Detection Algorithms. 2022, Comput. Sys. Appl., 31(7), 35–45.

[5] Lyu, L., Cheng, H., & Zhu, H. Progress of Research and Application of Object Detection Based on Deep Learning. 2022, Elec. Pack., 22(1): 010307.

[6] Geng, C., Song, P., & Cao, L. Research Progress of YOLO Algorithm in Target Detection. 2022, J. Ord. Equip. Eng.

[7] Nguyen, D., Hong, H., Kim, K., & Park, K. Person recognition system based on a combination of body images from visible light and thermal cameras. 2017 Sensors, 17(3), 605.

[8] YOLO Algorithm for Object Detection Explained [+Examples]. https://www.v7labs.com/blog/yolo-object-detection