FRI-2G.302-1-CCT2-02

EXPLORING OBJECT DETECTION ALGORITHMS: A COMPREHENSIVE OVERVIEW AND COMPARATIVE STUDY [15]

Assist. Prof. Georgi Georgiev, PhD Student
Department of Telecommunications,
"Angel Kanchev" University of Ruse
Tel.: 00359 82 888 353
E-mail: gdgeorgiev@uni-ruse.bg

Prof. Georgi Hristov, PhD
Department of Telecommunications,
"Angel Kanchev" University of Ruse
Tel.: 00359 82 888 663
E-mail: ghristov@uni-ruse.bg

Assoc. Prof. Plamen Zahariev, PhD
Department of Telecommunications,
"Angel Kanchev" University of Ruse
Tel.: 00359 82 888 663
E-mail: pzahariev@uni-ruse.bg

Assist. Prof. Diyana Kinaneva, PhD
Department of Telecommunications,
"Angel Kanchev" University of Ruse
Tel.: 00359 82 888 353
E-mail: dkyuchukova@uni-ruse.bg

Abstract: *Object detection is a fundamental computer vision task with diverse applications, ranging from autonomous driving to image and video analysis. Over the years, numerous algorithms have been developed to tackle this problem, each with its own set of strengths and weaknesses. In this paper, a comprehensive survey and comparative analysis of various object detection algorithms are presented, shedding light on their underlying principles, methodologies, and performance characteristics.*

This study covers a wide spectrum of object detection techniques, including traditional approaches like Haar cascades and template matching, as well as modern deep learning-based methods such as Faster R-CNN, YOLO, and SSD. The evolution of object detection algorithms is discussed, emphasizing the pivotal role of convolutional neural networks (CNNs) in revolutionizing the field.

Keywords: *Object Detection, Computer Vision, Deep Learning, Convolutional Neural Networks (CNNs)*

INTRODUCTION

In the rapidly evolving landscape of computer vision, object detection emerges as a pivotal technology with far-reaching implications and diverse applications. From enabling autonomous vehicles to navigate complex environments to refining the analysis of intricate image and video data, the significance of accurate and efficient object detection cannot be overstated. This paper aims to provide a comprehensive survey and a comparative analysis of the object detection algorithms that have been developed over the years, each contributing uniquely to this dynamic field.

[15] Докладът е представен на научната сесия на 27.10.2023 в секция „Комуникационна и компютърна техника" с оригинално заглавие на български език: ИЗСЛЕДВАНЕ НА АЛГОРИТМИ ЗА ОТКРИВАНЕ НА ОБЕКТИ: ЦЯЛОСТЕН ПРЕГЛЕД И СРАВНИТЕЛНО ИЗСЛЕДВАНЕ

The evolution of object detection has been marked by significant milestones, from the early days of traditional methods like Haar cascades and template matching to the current era dominated by advanced deep learning techniques. These methodologies have progressively reshaped the landscape of object detection, offering improvements in accuracy, speed, and reliability. The advent of convolutional neural networks (CNNs), in particular, has been a game-changer, revolutionizing the approach to object detection and setting new standards in performance.

This study endeavors to explore the wide spectrum of object detection algorithms, providing insights into their underlying principles, methodologies, and practical implications. By examining both the traditional approaches and the modern deep learning-based methods such as Faster R-CNN, YOLO, and SSD, this paper highlights the evolution and current state of object detection technologies. Through a detailed comparative analysis, the strengths, limitations, and performance characteristics of these varied techniques are sought to be illuminated, offering a clear perspective on the respective roles and effectiveness of each method in the field of computer vision.

TRADITIONAL OBJECT DETECTION TECHNIQUES

Traditional object detection techniques, which predate the deep learning era, played a crucial role in the early development of computer vision. These techniques, while less sophisticated than their modern counterparts, laid the foundation for the field and are still relevant in certain applications.

One of the earliest and most widely used methods in traditional object detection is the Haar cascade classifier. Developed in the early 2000s, it is particularly effective for face detection. The Haar cascade method relies on Haar-like features, which are simple contrast features used to identify objects in an image. These features are then used to train a cascade of classifiers, which can quickly discard non-relevant regions in an image, focusing on areas likely to contain the object of interest (Soo, S., 2014). This approach is efficient but often limited in its ability to handle variations in scale, orientation, and lighting.

Another traditional technique is template matching. This method involves sliding a template image across the target image and computing a similarity measure at each position (Banharnsakun, A., & Tanathong, S., 2014). The highest similarity scores indicate the likely positions of the object. Template matching is straightforward and effective for detecting objects with little variation in appearance but struggles with scale and rotational changes.

Histogram of Oriented Gradients (HOG) is another notable technique, especially popular for pedestrian detection. It involves dividing the image into small regions and calculating the distribution (histogram) of edge orientations within each region. These histograms are then used as features for object detection (Ren, H., & Li, Z. N., 2014).

While these traditional methods have been largely overshadowed by the advent of deep learning, they are still valuable in scenarios where computational resources are limited or when working with simpler, well-defined visual tasks. The principles and techniques developed in this era continue to influence and inform current research in object detection.

ROLE OF DATA AND PERFORMANCE METRICS FOR OBJECT DETECTION

The role of data in object detection and the performance metrics used to evaluate these systems are two critical aspects that significantly influence the effectiveness of object detection models.

In object detection, data primarily refers to the images or videos used to train, validate, and test the models. The quality, diversity, and volume of this data play a pivotal role in the development of robust and accurate object detection systems. Large and varied datasets allow models to learn and generalize better, making them more effective in real-world scenarios. These datasets need to be meticulously annotated, usually by bounding boxes around each object of interest, along with labels indicating the object's class. The process of data collection and annotation is both time-consuming and labor-intensive, but it is crucial for building effective models. Well-known datasets in object detection include PASCAL VOC, MS COCO, and ImageNet, each

offering a wide range of images with varied objects and scenarios, providing a comprehensive ground for training and evaluating object detection models.

Performance metrics in object detection are designed to evaluate how well a model can detect and correctly classify objects in images. Common metrics include precision, recall, and the F1 score. Precision measures the proportion of correctly identified positive instances among all instances identified as positive, while recall measures the proportion of actual positive instances that were correctly identified. The F1 score provides a balance between precision and recall, giving a single measure of the model's accuracy. Another crucial metric is the Intersection over Union (IoU), which assesses the accuracy of the bounding box by calculating the area of overlap between the predicted bounding box and the ground truth box, divided by the area of union of these two boxes. A higher IoU indicates a more accurate prediction.

Additionally, more comprehensive metrics like mean Average Precision (mAP) are often used, especially in benchmarking models on datasets like PASCAL VOC or MS COCO. The mAP takes into account the precision-recall curve for each class and computes the average, providing a single performance figure that considers both the accuracy of the bounding boxes and the correct classification of objects.

In summary, the role of data in object detection is fundamental, as it directly influences the model's ability to learn and perform accurately. Simultaneously, performance metrics provide a means to quantitatively assess and compare the effectiveness of different object detection models, ensuring a standard for evaluating advancements in the field.

THE ADVENT OF DEEP LEARNING IN OBJECT DETECTION

The advent of deep learning in object detection marks a significant milestone in the field of computer vision, representing a paradigm shift from traditional methods to more sophisticated, data-driven approaches. This transformative phase is characterized by the transition from relying on handcrafted features and heuristic rules to leveraging the power of neural networks that learn feature representations directly from data.

Deep learning, particularly through the use of Convolutional Neural Networks (CNNs), has redefined the possibilities in object detection.

Convolutional Neural Networks (CNNs) have become the backbone of modern object detection, representing a monumental shift in the way images are analyzed and interpreted. At their core, CNNs are a class of deep neural networks, uniquely structured to process data that comes in the form of arrays, making them especially suitable for image processing tasks.

The architecture of a CNN is distinct in its use of convolutional layers, which automatically and adaptively learn spatial hierarchies of features from input images. These features range from simple edges in the initial layers to more complex patterns in the deeper layers. A typical CNN architecture consists of a series of convolutional layers interspersed with activation functions like ReLU (Rectified Linear Unit), pooling layers, and fully connected layers towards the end.

The convolutional layers act as feature extractors. They apply a number of filters to the input, creating feature maps that represent the presence of specific features or patterns at different locations in the image. The pooling layers, usually following the convolutional layers, serve to reduce the spatial size of the representation, decreasing the amount of computation and weights in the network, and thereby helping to control overfitting.

One of the pivotal moments in the history of CNNs was the introduction of AlexNet in 2012. This deep CNN outperformed all previous methods in the ImageNet Large Scale Visual Recognition Challenge, a benchmark competition in image classification. AlexNet's success stemmed from its deep architecture, which was able to learn complex patterns in large-scale image data.

Following AlexNet, other sophisticated CNN architectures emerged, each introducing novel concepts and structures to improve performance. For instance, VGGNet, notable for its simplicity and depth, used small convolution filters to build deeper networks. GoogLeNet (or Inception) introduced inception modules, allowing the network to choose from different types of features at each level. ResNet, another significant architecture, brought in the concept of residual learning with

skip connections to enable the training of very deep networks without the problem of vanishing gradients.

The application of CNNs in object detection led to the development of various models that significantly advanced the field. Models like R-CNN, Fast R-CNN, and Faster R-CNN leveraged the power of CNNs not just for classifying images, but also for accurately locating and identifying multiple objects within a single image.

REGION-BASED CONVOLUTIONAL NEURAL NETWORKS (R-CNN) AND VARIANTS

Region-Based Convolutional Neural Networks (R-CNN) and their subsequent variants represent significant advancements in the field of object detection, leveraging the power of Convolutional Neural Networks (CNNs) to not only classify but also accurately locate objects within images.

The original R-CNN model, introduced by Ross Girshick et al., was a pioneering approach that combined high-capacity CNNs with region proposal methods for object detection. The key idea was to first generate potential object regions in an image – known as region proposals – and then use a CNN to extract features from each region independently, followed by a classifier to determine the object's class. This method marked a significant shift from traditional object detection techniques, offering much higher accuracy by effectively integrating region proposals with deep feature extraction.

However, R-CNN had its limitations, primarily in terms of computation efficiency. The model was slow due to the independent processing of each region proposal through the CNN, leading to a significant amount of redundant computation. To address this, the Fast R-CNN model was developed. Fast R-CNN improved upon the original design by passing the entire image through the CNN just once to create a feature map, from which features corresponding to each region proposal were extracted (Girshick, R., 2015). This approach, known as RoI (Region of Interest) Pooling, significantly reduced the computational load and improved the processing speed.

Building on Fast R-CNN, the Faster R-CNN model was introduced to further enhance the efficiency and speed of object detection. Faster R-CNN integrated the process of generating region proposals into the network itself, introducing a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network. This shared feature extraction between the RPN and the detection network streamlined the process, enabling nearly real-time object detection.

Faster R-CNN became a cornerstone in object detection due to its accuracy and speed. It was capable of processing images with complex scenes and multiple objects, making it suitable for a wide range of applications. The integration of the RPN also meant that the network could be trained end-to-end, which was a significant improvement over the multi-stage training required by its predecessors.

Subsequent variants and improvements continued to build on the foundation laid by R-CNN. For instance, models like Mask R-CNN extended Faster R-CNN by adding a branch for segmenting objects at the pixel level, further increasing the versatility and accuracy of the model.

SINGLE SHOT DETECTORS (SSD) AND YOU ONLY LOOK ONCE (YOLO)

Single Shot Detectors (SSD) and You Only Look Once (YOLO) represent two groundbreaking approaches in the field of object detection, known for their speed and efficiency, particularly in real-time applications. These models marked a significant departure from the earlier, region proposal-based methods like R-CNN and its variants, streamlining the object detection process by eliminating the need for a separate region proposal step.

YOLO, as its name suggests, processes an entire image in a single evaluation, making it one of the fastest object detection algorithms available. Developed by Joseph Redmon et al., YOLO divides the input image into a grid, and each grid cell predicts a certain number of bounding boxes along with confidence scores for those boxes (Liu, C., Tao, Y., Liang, J., Li, K., & Chen, Y., 2018). These confidence scores reflect how confident the model is that the box contains an object and also how accurate it thinks the box is compared to the ground truth. Additionally, each box predicts the class of the object contained within it. One of the key strengths of YOLO is its ability to look at the

entire image during training and testing, which helps it understand the context of objects, reducing false positives significantly.

Following the initial version, subsequent iterations of YOLO were released, each improving upon the accuracy and speed. For instance, YOLOv2 and YOLOv3 incorporated various enhancements, including using anchor boxes to improve bounding box prediction, and applying multi-scale training methods to increase detection accuracy across different object sizes.

On the other hand, SSD, introduced by Wei Liu et al., takes a different approach. It operates by discretizing the output space of bounding boxes into a set of default boxes over different aspect ratios and scales per feature map location. During prediction, the model generates scores for the presence of each object category in each default box and produces adjustments to the box to better match the object shape. Moreover, SSD performs these detections over multiple scales by operating on feature maps of different sizes. This multi-scale approach allows SSD to detect objects of various sizes effectively (Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., & Berg, A.C., 2016).

SSD's main advantage lies in its balance between speed and accuracy. While it may not match the speed of YOLO, it often achieves higher detection accuracy, particularly for small objects, due to its multi-scale detection strategy.

Both YOLO and SSD have significantly influenced the landscape of object detection. Their ability to perform detections in a single shot, without the need for a separate region proposal and refinement steps, makes them highly efficient and well-suited for real-time applications. These models exemplify the ongoing advancements in deep learning-based object detection, demonstrating a continual push towards faster, more efficient, and more accurate methods.

COMPARISON

This section will compare Faster R-CNN, YOLO and SSD across multiple dimensions – mean Average Precision (mAP), Frames Per Second (FPS) and batch size (Table 1). Through this analysis, the trade-offs of each algorithm will be explored and their suitability for various applications will be discussed, providing insights into the selection of the most appropriate model based on specific performance criteria.

The MS COCO 2017 dataset was utilized as the evaluation platform, providing a varied array of images and object classes. Recognized widely as a benchmarking standard, it delivered an extensive array of real-world scenarios, from urban congestion to diverse natural environments, ensuring a rigorous examination of each model's ability to detect objects.

Table 1. Results from comparison

Algorithm	mAP (%)	FPS (Batch Size = 1)	FPS (Batch Size = 8)
Faster R-CNN (VGG16)	73.2	7	Not Applicable
Fast YOLO	52.7	155	Not Applicable
YOLO (VGG16)	66.4	21	Not Applicable
SSD300	74.3	46	59
SSD512	76.8	19	22

Faster R-CNN (VGG16) has shown impressive accuracy with a mAP of 73.2 %. This performance comes at the cost of speed, achieving only 7 FPS. With its architecture relying on a two-stage process – wherein the first stage proposes regions and the second stage classifies them – the algorithm is computationally intensive, making it less ideal for real-time applications. However, its high mAP makes it suitable for applications where accuracy is more critical than speed.

Fast YOLO is the fastest among the algorithms, running at 155 FPS, but its accuracy is significantly lower, with a mAP of 52.7 %. YOLO's architecture allows it to predict both bounding boxes and class probabilities in one evaluation of the network, making it incredibly fast. However, this speed comes at the cost of precision, particularly in detecting small objects or objects with a lot of overlap.

YOLO (VGG16) strikes a balance between the first two, with a mAP of 66.4 % and a speed of 21 FPS. It shows that adjustments and improvements in the YOLO architecture, potentially in feature extractor choice and training techniques, can yield a better trade-off between speed and accuracy.

SSD comes in two variants based on input resolution: SSD300 and SSD512. Both variants outperform YOLO in accuracy while maintaining a high frame rate. SSD300 achieves a mAP of 74.3 % at 46 FPS, whereas SSD512 achieves a mAP of 76.8 % at a slower 19 FPS due to the higher resolution input. The SSD models, with their different input resolutions, demonstrate how increasing the resolution can improve accuracy, as SSD512 outperforms SSD300, albeit at a reduced frame rate.

When batch sizes are increased from 1 to 8, SSD300's FPS improves from 46 to 59, and SSD512's from 19 to 22, illustrating how batch processing can improve throughput. However, it is worth noting that batch processing is not always feasible in real-time scenarios where latency is a concern.

CONCLUSION

Traditional techniques like Haar cascades and template matching laid the groundwork, showing the potential of automated systems in identifying objects within an image. The baton was then passed to deep learning methods, which have revolutionized the field. Convolutional Neural Networks (CNNs) emerged as a powerful tool for feature extraction, leading to the development of advanced algorithms like Faster R-CNN, YOLO, and SSD. These models have pushed the boundaries of accuracy and speed, allowing for real-time applications and opening up new possibilities for innovation across various industries.

The choice among Faster R-CNN, YOLO, and SSD models depends on the specific requirements of the application at hand. Faster R-CNN is preferable for high-accuracy needs, Fast YOLO for scenarios requiring speed, and SSD for a balanced approach. The trade-offs between speed and accuracy and the impact of input resolution and batch size processing are critical considerations in selecting the appropriate model for a given object detection task.

In conclusion, object detection stands as a testament to the remarkable progress in computer vision. As the field continues to advance, it holds the promise of even more sophisticated applications, making the interaction between humans and machines more seamless and intuitive. The future of object detection is bright and brimming with potential, continuing to evolve as a cornerstone of artificial intelligence.

ACKNOWLEDGMENT

This work was supported by the Bulgarian Ministry of Education and Science under the National Research Programme “Smart Agriculture” approved by Decision of the Ministry Council №866/26.11.2020.

The work presented in this paper is completed with the support of Project 2023 – FEEA – 03 “Simulation and experimental study on the methods and mechanisms for data confidentiality and data integrity in the modern local and wireless communication networks”, financed under the Scientific and Research Fund of the University of Ruse “Angel Kanchev”.

REFERENCES

Banharnsakun, A., & Tanathong, S. (2014). *Object Detection Based on Template Matching through Use of Best-So-Far ABC.* Computational intelligence and neuroscience, vol. 2014, 8 pages, https://doi.org/10.1155/2014/919406.

Girshick, R. (2015). *Fast R-CNN*. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

Liu, C., Tao, Y., Liang, J., Li, K., & Chen, Y. (2018). *Object Detection Based on YOLO Network.* In 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC), pp. 799-803.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., & Berg, A. C. (2016). *SSD: Single Shot MultiBox Detector*. In Proceedings of the European Conference on Computer Vision (ECCV), Computer Vision – ECCV 2016. https://doi.org/10.1007/978-3-319-46448-0_2.

Ren, H., & Li, Z. N. (2014). *Object Detection Using Edge Histogram of Oriented Gradient.* In 2014 IEEE International Conference on Image Processing (ICIP), pp. 4057-4061.

Soo, S. (2014). Object Detection Using Haar-Cascade Classifier. Institute of Computer Science, University of Tartu, 2(3), pp. 1-12.