Intermediate Project Report for CV Project 2

Devansh Bansal
B20CS094
bansal.11@iitj.ac.in

Abstract

In this report, I am presenting an overview of the problem of Object Detection, along with insights from initial solutions of the problem, till some of the solutions found recently. This report includes solutions for special cases like faces and cars [1], but doesn't restrict to only special cases. It also includes the quite famous initial solution using Boosted Cascade [2]. After such a radical solution, there were many variants of the proposed solution, but without any significant improvement, as all these solutions were of pre-deep learning era. After the advent of Deep-Learning, some novel solutions were proposed for the problem in discussion and among them, one of the quite impactful solutions, R-CNN [3] is discussed here. Also to show that, changing the architecture or method isn't always the only path for better performance, a novel Loss-Function called Focal Loss is discussed here [4]. The contributions of all these solutions to the problem and the community is briefly discussed along with the future plan of implementation of these methods.

1. Introduction

Object detection is a crucial issue in computer vision that entails locating and identifying objects in images or videos. Variations in object aspect, size, and occlusion, among other factors, make it a difficult task. Object detection is a crucial component of numerous practical applications, such as autonomous driving, face recognition, robotics, and surveillance. In recent years, researchers have paid a great deal of attention to this issue, resulting in the development of numerous solutions.

Object detection aims to precisely identify the presence and location of objects in images or videos. Typically, it involves two primary tasks: object localization and classification. Object localization is the process of determining the location of objects in an image, which is typically depicted by a bounding box. Assigning a label to each object, such as a vehicle, a pedestrian, or a traffic light, is part of object classification. Object detection is a difficult endeavor because it requires the ability to detect objects at various dimensions, orientations, and perspectives. Additionally, it must accurately differentiate between distinct object categories.

This report discusses four papers proposing unique object detection methods. The first paper, titled "A Statistical Method for 3D Object Detection Applied to Faces and Cars," [1] proposes a statistical method for 3D object detection applicable to faces and cars. The authors present a probabilistic model for estimating the position and orientation of three-dimensional objects. They use the Expectation-Maximization algorithm to estimate the model's parameters and demonstrate that it can obtain accurate face and car detection results.

The second paper, "Rapid Object Detection Using a Boosted Cascade of Simple Features," [2] presents a boosted cascade of simple features that facilitates rapid object detection. The authors propose a novel algorithm for selecting a small subset of the most discriminative features for object detection. They use these features to train a cascade of classifiers capable of detecting objects in images quickly. The authors demonstrate that their method is effective and achieves cutting-edge detection results.

The third paper, "Rich feature hierarchies for accurate object detection and semantic segmentation," [3] proposes a deep learning-based method that makes use of rich feature hierarchies to enhance the accuracy of object detection and semantic segmentation. The authors suggest a deep convolutional neural network capable of learning hierarchical image representations. They utilize these representations to simultaneously conduct object detection and semantic segmentation. The authors demonstrate that their method obtains state-of-the-art performance on multiple object detection benchmarks.

The final paper, titled "Focal Loss for Dense Object Detection," [4] presents focal loss, a novel loss function that addresses the imbalance between foreground and background samples in dense object detection. The authors propose a modified variant of the cross-entropy loss that assigns greater training weights to difficult examples. They demonstrate that their method enhances the detection pre-

cision of dense objects, which are extremely difficult to detect.

Overall, these publications showcase the variety of approaches that have been put out for object detection and show how work is still being done to improve methods and make them more effective. In the first paper, a statistical technique is suggested, in the second, a boosted cascade of basic characteristics, in the third, a deep learning-based strategy, and in the fourth, a unique loss function. This also shows the varied amount of methodologies applicable for a single problem, and hence the vastness of the problem can be understood from these quite-a-lot varied solutions.

We will go into great length on the strengths and weaknesses of each work in the sections that follow, as well as how those papers fit into the larger context of object detection research.

1.1. Contributions

I am doing this project single-handedly, so all of its work, analysis, inferences, results are done by me.

1.2. A Statistical Method for 3D Object Detection Applied to Faces and Cars

In this paper, a statistical method for 3D object detection is suggested. The position and orientation of objects in 3D space are estimated by the authors using a probabilistic approach. Using an Expectation-Maximization technique, the model calculates the properties of the objects, including their 3D location and orientation. The approach is novel as it does not require prior information about the shape and size of objects, making it more flexible and efficient.

The initialization stage, the parameter estimate step, and the object identification step are the three key steps of the suggested methodology. Using a clustering approach, the picture is separated into areas of interest in the initialization stage. The approach uses an Expectation-Maximization algorithm in the parameter estimation stage to estimate object parameters like 3D location and orientation i.e. by maximizing the likelihood of detecting the picture data given the 3D object model and the camera parameters. The likelihood of viewing the image data given the 3D object model and the camera parameters is compared to the chance of detecting the same image data given the null hypothesis that no object is present in the image. The probability ratio is the measurement used to assess whether an object is visible in the image.

The authors add a threshold to the likelihood ratio to strengthen and improve the identification procedure. The object is regarded as being present in the image if the likelihood ratio is higher than the threshold; otherwise, it is seen as being missing. The false positive rate, which is the likelihood of identifying an object when none is present, and the false negative rate, which is the likelihood of missing an object when one is present, are used to determine the threshold. The authors demonstrate how the threshold can be set to have a low false positive rate and a high true positive rate, which is the likelihood of spotting an object when one is there.

The Yale Face Dataset and the MIT-Car dataset were two datasets that the authors used to test their technology and compare it to existing 3D object identification techniques. With a false-positive rate of less than 1%, they claimed that their method produced accurate detection findings. The outcomes showed that the suggested strategy surpassed existing approaches in terms of speed and accuracy and was computationally efficient.

1.3. Rapid Object Detection using a Boosted Cascade of Simple Features

This seminal paper proposed a quick and effective approach for identifying things in photos. The technique—known as the Viola-Jones algorithm—uses a boosted cascade of straightforward characteristics to rapidly weed out areas of a picture that are most likely to contain the item of interest and so lessen the computing burden of the identification process.

The technique entails training a series of classifiers, each of which consists of a weak classifier, such as a decision tree, and a weak feature, such as a Haar-like feature. Each classifier in the cascade is trained on a subset of the difficult instances that it misclassified in the previous step, together with the negative examples, in order to reduce false positives at each stage. In order to lower the computing cost of the approach, the authors offer a unique feature selection algorithm that enables them to choose a limited number of highly discriminative features.

The MIT+CMU face database was one of the datasets used by the authors to test their methodology, and they found that it had a 95% detection rate and a 50% false positive rate. The outcomes demonstrated how much faster the new approach was than the previous methods, making it appropriate for real-time applications.

1.4. Rich feature hierarchies for accurate object detection and semantic segmentation

In order to improve upon the pre-Deep-Learning Methods, this paper suggests a deep learning-based approach for object detection and semantic segmentation. The technique, referred to as the Region-based Convolutional Neural Network (R-CNN), entails instructing a deep neural network on a sizable dataset of pictures in order to acquire a hierarchical collection of features that may be applied to object recognition and semantic segmentation.

Region proposal, feature extraction, and object categorization are the method's three primary phases. Using a selective search technique, the region suggestion stage of the

approach creates a list of potential object-containing regions in the picture. Using a deep convolutional neural network, the feature extraction stage of the technique retrieves a collection of features for each candidate region. The technique employs a support vector machine to determine if each candidate region contains an object or not, as well as to forecast the object's class if it does.

The PASCAL VOC and MS COCO datasets, among others, were used to test the authors' method, and they found that it produced state-of-the-art results in tasks requiring object recognition and semantic segmentation. The findings demonstrated that the suggested method was noticeably more accurate than the alternatives, making it a potential strategy for practical applications.

1.5. Focal Loss for Dense Object Detection

For training deep neural networks for dense object recognition tasks, this paper suggests a novel loss function termed focal loss. In dense object detection tasks, where there are frequently many more background instances than positive examples, the authors discovered the class imbalance problem. This issue is solved by the focused loss function, which reduces the loss applied to correctly categorized instances so that the network may concentrate more on challenging cases.

A modified cross-entropy loss called the focal loss function is described as having a modulating factor dependent on the expected probability of the right class. The modifying factor rises as the likelihood of the proper class falls, emphasising challenging cases. The authors showed that dense object identification tasks, such as object detection in natural pictures and instance segmentation in biological images, considerably benefit from the focus loss function.

This study introduces the cutting-edge object detection model RetinaNet, which trains with a focal loss function. RetinaNet uses a feature pyramid network (FPN) to recognise objects at various sizes in order to solve the issue of scale variation in object detection tasks. The FPN is a multi-scale feature extraction network that creates a feature pyramid by combining features from several backbone network levels utilising lateral connections.

For object identification, RetinaNet employs a two-branch architecture, with one branch used for classification and the other for regression. While the regression branch forecasts the bounding box coordinates of the item, the classification branch forecasts the likelihood that an object will belong to a specific class. Both branches are trained using the focal loss function, with the classification branch's projected probability of the right class serving as the modulating factor.

2. Current Progress and Future Work

I thoroughly studied the aforementioned publications, comprehended their methodology and findings, performed their analyses, and learned which was superior. Only these four papers were chosen because of the path they paved from the first special examples to generic ones and subsequently, with the birth of DL, how the problem was resolved in a very different way. Although Computer Vision and Deep Learning are thought to be extremely distinct, they may work together and are almost equally significant for issues and solutions in the real world. My current goal is to duplicate at least two of the four publications mentioned above in their entirety, perform a thorough analysis, and then compare the datasets utilized, the conclusions obtained, and the execution time. Since our environment is resource-constrained, I would try to replicate the methods on the datasets they were presented and tested on, but if I ran out of time or resources, I would do the same tasks on a specific subset of the dataset that was used, and that subset would be selected at random while maintaining the overall distribution of the whole dataset.

References

[1] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, page 1746, Los Alamitos, CA, USA, jun 2000. IEEE Computer Society. 1

[2] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, page 511, Los Alamitos, CA, USA, dec 2001. IEEE Computer Society. 1

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, Los Alamitos, CA, USA, jun 2014. IEEE Computer Society. 1

[4] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, Los Alamitos, CA, USA, oct 2017. IEEE Computer Society. 1