THE TECHNIQUE OF REAL TIME OBJECT TRACKING

Prof. Shivaji Goroba Shinde1 ; Mr. Shubham Suresh Patil

Shivajishinde261@gmail.com ; shubhpatil2211@gmail.com

Dr.Babasaheb Ambedkar Technological University,Lonere ,Dist.Raigad

TPCT's College of Engineering , Osmanabad (Dharashiv), India.

Abstract

ABSTRACT Modern world is drenched with an enormous amount of visual data. Object tracking is the keystone of many computer vision applications. Identifying objects from an image or a video sequence is a primary and demanding task in today's world. The purpose of visual object tracking in successive video frames is to detect or connect target objects. It has many applications in the real world, including surveillance, face detection medical image processing, traffic control and analysis, and many more. While many approaches and breakthroughs in this field have led to the evolution of a huge set of unique algorithms, it remains a very stimulating problem. Due to the ginormous set of environmental and other factors, it is next to impossible to propose a global tracking algorithm. However, the selecting the most suitable algorithm does not depend only on the concept of the algorithm but also on its implementation. In this paper, we try to use YOLO or "You Only Look Once" with several configurations of the moving image feature to recognize objects. Additionally, we suggest a system that, by adaptively regulating the cycle of object detection and tracking, can provide real-time performance in different edge computing contexts.

Keywords: Object, detection, tracking, Artificial intelligence, Open CV, python

INTRODUCTION

Object detection and image recognition are related but distinct computer vision techniques. Object detection goes beyond image recognition by not only identifying objects in an image but also predicting their locations with bounding boxes. Object detection algorithms use machine learning or deep learning techniques to analyze images or videos and identify objects of interest, and then draw bounding boxes around them to indicate their precise locations. These bounding boxes are often accompanied by labels that describe the objects detected, providing more detailed information about the objects in the image compared to image recognition, which simply assigns labels to entire images without identifying specific objects or their locations. Object detection is widely used in various applications such as autonomous vehicles, surveillance systems, facial recognition, and augmented reality, among others.

Traditional machine learning-based methods for object detection often involve handcrafted feature engineering, where various aspects of an image, such as color histograms or edges, are extracted as features, and then these features are used as input into a separate regression model to predict the location of objects and their labels. These methods can work well in certain scenarios but may require manual tuning and are limited by the effectiveness of the handcrafted features. In contrast, deep learning-based techniques, such as convolutional neural networks (CNNs), have emerged as state-of-the-art approaches for object detection. CNNs are capable of automatically learning relevant features from raw image data during the training process, eliminating the need for handcrafted features. CNNs can learn complex patterns and representations from large amounts of labeled data, making them highly effective in detecting objects with high accuracy.

Deep learning-based object detection techniques, such as region-based CNNs (R-CNN), You Only Look Once (YOLO), and Single Shot MultiBox Detector (SSD), have become state-of-the-art methods for accurate and real-time object detection in various applications, such as autonomous driving, surveillance, and image retrieval systems, among others. They offer higher accuracy and faster processing speeds compared to traditional ML-based methods, making them the preferred choice in many practical scenarios.

What is an image?

Images can be captured and saved using a variety of media, including pictures, paintings, drawings, and digital data. Images are a visual depiction of an object or scene. An picture is often represented in digital form as a rectangular array of pixels, each of which has a unique colour or grayscale value that affects how the image looks as a whole. An picture's resolution is determined by how many pixels make up the image; higher resolution images have more pixels and, consequently, more detail.

GRAY SCALE IMAGE:

An image that is purely grayscale has shades of gray, ranging from black to white, as its only colors. Each pixel in a digital grayscale image is represented by a single brightness value, with lower values denoting darker shades and higher values denoting lighter shades. Since grayscale images may be saved and processed more effectively than color images while still transmitting crucial information about the intensity or brightness of the underlying data, they are frequently utilized in a variety of industries, including medical imaging, computer graphics, and image processing.

COLOR IMAGE:

As opposed to a grayscale image, which just contains different shades of gray, a color image is one that contains colors. Each pixel in a digital color image has numerous values that represent the intensities of the image's primary hues, which are commonly red, green, and blue. (RGB). Because they offer more visual information than grayscale images and are more suited to representing real-world landscapes and objects, color images are frequently utilized in a variety of industries, including photography, computer graphics, and medical imaging. Color images are valuable in a variety of image processing and computer vision applications because they may be utilized to extract additional information through color-based segmentation or feature extraction.

Method

There are two main differnet methods are use in this detection methods

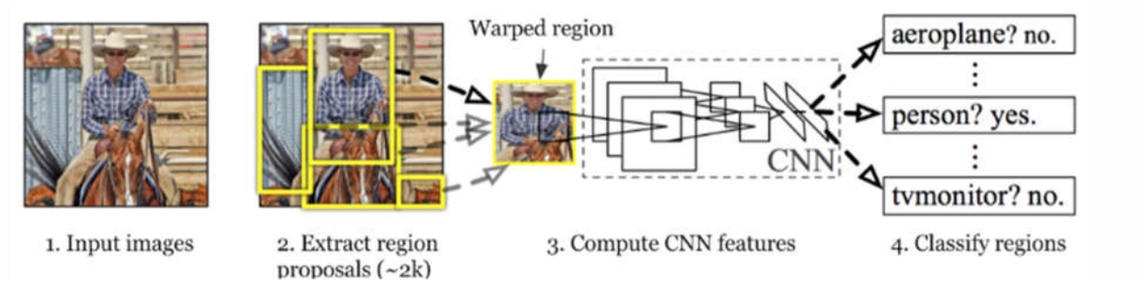
1. R-CNN

2.yolo

3.SSD

1. R-CNN

Since Convolution Neural Network (CNN) with a fully connected layer is not able to deal with the frequency of occurrence and multi objects. So, one way could be that we use a sliding window brute force search to select a region and apply the CNN model to that, but the problem with this approach is that the same object can be represented in an image with different sizes and different aspect ratios. While considering these factors we have a lot of region proposals and if we apply deep learning (CNN) to all those regions that would computationally very expensive



Ross Girshick et al in 2013 proposed an architecture called R-CNN (Region-based CNN) to deal with this challenge of object detection. This R-CNN architecture uses the selective search algorithm that generates approximately *2000* region proposals. These *2000* region proposals are then provided to CNN architecture that computes CNN features. These features are then passed in an SVM model to classify the object present in the region proposal. An extra step is to perform a bounding box regressor to localize the objects present in the image more precisely.

Challenges of R-CNN

- The selective Search algorithm is very rigid and there is no learning happening in that. This sometimes leads to bad region proposal generation for object detection.
- Since there are approximately *2000* candidate proposals. It takes a lot of time to train the network. Also, we need to train multiple steps separately (CNN architecture, SVM model, bounding box regressor). So, This makes it very slow to implement.
- R-CNN can not be used in real-time because it takes approximately *50 sec* to test an image with a bounding box regressor.
- Since we need to save feature maps of all the region proposals. It also increases the amount of disk memory required during training.

2. yolo

Humans glance at an image and instantly know what objects are in the image, where they are, and how they interact. The human visual system is fast and accurate, allowing us to perform complex tasks like driving with little conscious thought. Fast, accurate algorithms for object detection would allow computers to drive cars without specialized sensors, enable assistive devices to convey real-time scene information to human users, and unlock the potential for general purpose, responsive robotic systems.

YOLO is refreshingly simple: see Figure 1. A single convolutional network simultaneously predicts multiple bounding boxes and class probabilities for those boxes. YOLO trains on full images and directly optimizes detection performance. This unified model has several benefits over traditional methods of object detection. First, YOLO is extremely fast. Since we frame detection as a regression problem we don't need a complex pipeline. We simply run our neural network on a new image at test time to predict detections. Our base network runs at 45 frames per second with no batch processing on a Titan X GPU and a fast version runs at more than 150 fps. This means we can process streaming video in real-time with less than 25 milliseconds of latency. Furthermore, YOLO achieves more than twice the mean average precision of other real-time systems

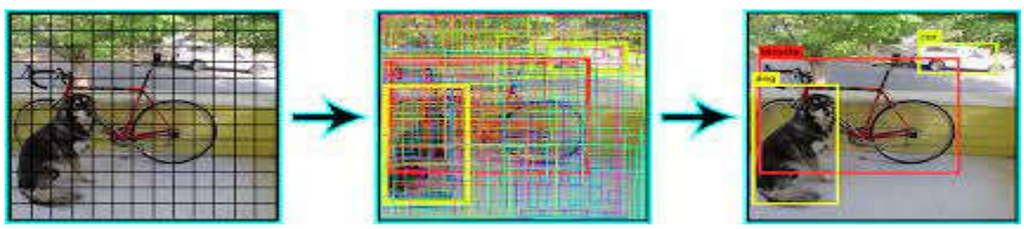


Figure 2: Example of YOLO (You Look At Once)

YOLO reasons globally about the image when making predictions. Unlike sliding window and region proposal-based techniques, YOLO sees the entire image during training and test time so it implicitly encodes contextual information about classes

as well as their appearance. Fast R-CNN, a top detection method [14], mistakes background patches in an image for objects because it can't see the larger context. YOLO makes less than half the number of background errors compared to Fast R-CNN.

3.SSD

SSD only needs an input image and ground truth boxes for each object during training. In a convolutional fashion, we evaluate a small set (e.g. 4) of default boxes of different aspect ratios at each location in several feature maps with different scales (e.g. 8 × 8 and 4 × 4 in (b) and (c)). For each default box, we predict both the shape offsets and the confidences for all object categories ((c1, c2, · · · , cp)). At training time, we first match these default boxes to the ground truth boxes. For example, we have matched two default boxes with the cat and one with the dog, which are treated as positives and the rest as negatives. The model loss is a weighted sum between localization loss.

The SSD approach is based on a feed-forward convolutional network that produces a fixed-size collection of bounding boxes and scores for the presence of object class instances in those boxes, followed by a non-maximum suppression step to produce the final detections. The early network layers are based on a standard architecture used for high quality image classification (truncated before any classification layers), which we will call the base network.

We then add auxiliary structure to the network to produce detections with the following key features: Multi-scale feature maps for detection We add convolutional feature layers to the end of the truncated base network. These layers decrease in size progressively and allow predictions of detections at multiple scales

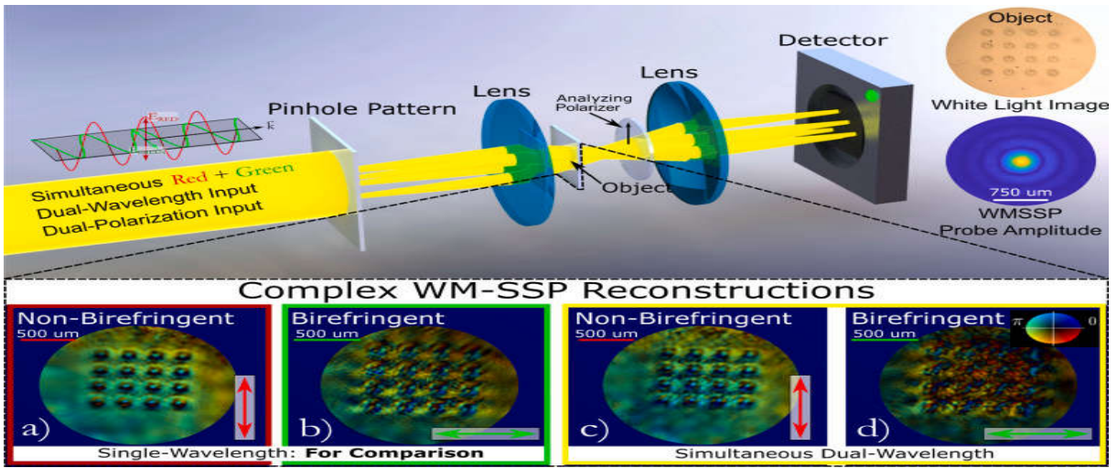


Figure 3 : Example of Single Shot Multibox Detector

The key difference between training SSD and training a typical detector that uses region proposals, is that ground truth information needs to be assigned to specific outputs in the fixed set of detector outputs. Some version of this is also required for training in YOLO[5] and for the region proposal stage of Faster R-CNN[2] and MultiBox[7]. Once this assignment is determined, the loss function and back propagation are applied end to-end. Training also involves choosing the set of default boxes and scales for detection as well as the hard negative mining and data augmentation strategies.

CONCLUSION

CONCLUSION: In order to accomplish object recognition and tracking, this project work proposes a straightforward, reliable methodology that can be used with object detection and tracking. This technique makes use of the YOLO Algorithm, which can anticipate and categorize bounding boxes in a single forward pass. This approach can purposefully improve the performance of detection. Compared to traditional machine learning algorithms, it is substantially faster. On the basis of accuracy, robustness, and computational effectiveness, the algorithms are evaluated. In this project, a comparative analysis of various object-identification techniques including RCNN, Faster RCNN, and YOLO is conducted. We discovered that YOLO is a lot quicker and more accurate than other object detection techniques. Thus, we trained the algorithm using the datasets. Real-time detection and tracking of the objects are possible with the YOLO algorithm. The camera module, which is a device that can be linked to a computer or desktop, provides the necessary input image. The PC or desktop itself allows us to see the outcomes. This technology could be applied and used in a variety of fields, including traffic analysis, face detection, medical image processing, and security monitoring.

FUTURE SCOPE

Object tracking is being more widely adopted by corporations, with uses ranging from personal security to workplace productivity. Object tracking is used across a wide range of image processing applications, including image retrieval, security, surveillance, automated driving systems, and machine analysis. There are still significant challenges in the realm of object detection. Regarding prospective outcomes for future use cases of object tracking, the possibilities are incalculable. Applications for object tracking include traffic analysis, surveillance and security, video correspondence, robot vision, and activity. Counting people can also be done using object detection. It is used to analyze festival crowd measurements or retail performance. They will typically become more challenging when people leave the picture quickly, likewise because individuals are no inflexible objects).

Person detection is a pivotal and required task in any intelligent video surveillance system because it provides the information needed to understand the semantics of the video recordings. 57 Due to the potential for enhancing security frameworks, it has a noticeable augmentation to automotive applications. Human detection is a task that computer vision frameworks undertake for locating and tracking people. Finding every instance of a person in a photograph is the challenge of person detection, which has most commonly been accomplished by scanning the entire image at all possible scales and comparing a small area at each scale with the known arrangements of people.

REFERENCES

[1] Mohana, HV Ravish Aradhya, "Object Detection and Tracking using Deep Learning and Artificial Intelligence for Video Surveillance Applications," (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 12, 2019, pp. 517-530.

[2] V. D. Nguyen et all., "Learning Framework for Robust Obstacle Detection, Recognition, and Tracking", IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 6, pp. 1633-1646, June 2017.

[3] P. Wang et all., "Detection of unwanted traffic congestion based on existing surveillance system using in freeway via a CNN-architecture traffic net", IEEE Conference on Industrial Electronics and Applications (ICIEA), Wuhan, 2018, pp. 1134-1139 .

[4] H. C. Baykara et all., "Real-Time Detection, Tracking and Classification of Multiple Moving Objects in UAV Videos", 29th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Boston, MA, 2017, pp. 945-950.

[5] K. Muhammad et all., "Convolutional Neural Networks Based Fire Detection in Surveillance Videos", IEEE Access, vol. 6, pp. 18174- 18183, 2018.

[6] Redmon, Joseph, et al." You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern Recognition', 2016.

[7] Aloysius, Neena, and M. Geetha." A review on deep convolutional neural networks." 2017 International .

[8] Z. Jiang, L. Zhao, S. Li and Y. Jia, "Real-time object detection method based on improved YOLOv4-tiny, " ArXiv, vol: abs/2011.04244, 2020.

[9] K. M. Babu and M. V. Raghunadh, "Vehicle number plate detection and recognition using bounding box method," 2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), 2016, pp. 106-110, Doi: 10.1109/ICACCCT.2016.7831610. 59

[10] K. V. Arya, S. Tiwari and S. Behwalc, "Real-time vehicle detection and tracking," 2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2016, pp. 1-6, Doi: 10.1109/ECTICon.2016.7561327.

[11] M. A. Bin Zuraimi and F. H. Kamaru Zaman, "Vehicle Detection and Tracking using YOLO and Deep SORT," 2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), 2021, pp. 23-29, Doi: 10.1109/ISCAIE51753.2021.9431784.

[12] X. Gu, Z. Chen, T. Ma, F. Li and L. Yan, "Real-Time vehicle detection and tracking using deep neural networks," 2016 13th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2016, pp. 167-170, Doi: 10.1109/ICCWAMTIP.2016.8079830.

[13] Z.Q. Zhao, P. Zheng, S.T. Xu, and X. Wu, "Object detection with deep learning,": A review. arXiv e-prints, arXiv:1807.05511, 2018.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection, " In 2016 IEEE conference on computer vision and pattern recognition, doi.org/10.1109/cvpr.2016.91 (pp. 779–788): IEEE.

[15] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6517-6525, Doi: 10.1109/CVPR.2017.690. [16] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” 2018. arXiv preprint arXiv:1804.02767.