YOLOV3-BASED OBJECT DETECTION FOR VISUALLY IMPAIRED

Manjunath H[1], Devadath Varma[2], MJ Shashank[3], Aadhar Kumar Mishra[4]

[1] *Dept of CS&E, Bangalore Institute of Technology, Bangalore, India*
E-mail: manjunathh@bit-bangalore.edu.in
[2] *Dept of CS&E, Bangalore Institute of Technology, Bangalore, India*
E-mail: 1bi19cs048@bit-bangalore.edu.in
[3] *Dept of CS&E, Bangalore Institute of Technology, Bangalore, India*
E-mail: 1bi19cs080@bit-bangalore.edu.in
[4] *Dept of CS&E, Bangalore Institute of Technology, Bangalore, India*
E-mail: 1bi19cs003@bit-bangalore.edu.in

ABSTRACT

The YOLOv3 algorithm, or "You Only Look Once version 3," offers a groundbreaking approach to object detection with significant implications for aiding the visually impaired. This innovative deep learning model excels in real-time object detection, making it a valuable tool for the blind and visually impaired. YOLOv3 operates by processing an image through a series of convolutional layers, identifying and precisely locating objects. For the visually impaired, this technology can be leveraged through custom applications or wearable devices that provide real-time object recognition, alerting users to the presence and position of objects in their environment. It has the potential to enhance independence and safety by enabling users to navigate unfamiliar surroundings more confidently. The YOLOv3 algorithm's remarkable speed and accuracy, along with its accessibility features, make it a powerful tool for creating inclusive solutions for the visually impaired, ultimately improving their quality of life and mobility.

Keyword-*YOLOv3, Object detection, Blind, Deep learning model*

1. INTRODUCTION

The YOLOv3 algorithm, which stands for "You Only Look Once version 3," represents a significant breakthrough in the field of computer vision and has substantial implications for empowering individuals who are blind or visually impaired. It is more than just an object detection tool; it is a game-changer for those seeking to navigate the visual world with confidence and independence. YOLOv3's core strength lies in its ability to rapidly and accurately identify objects within images and video streams, making it an ideal candidate for real-time object recognition applications.

What sets YOLOv3 apart is its efficiency and speed. Unlike traditional object detection models that require multiple passes through an image, YOLOv3 processes the entire scene in a single forward pass, producing near-instantaneous results. For individuals with visual impairments, this speed can translate into valuable real-time information about their surroundings. By harnessing the power of YOLOv3, developers can create applications and devices that provide audio cues or tactile feedback, enabling users to identify objects, obstacles, or even read text as they move through their environment. This introduction delves into the technical underpinnings of YOLOv3 and explores the immense potential it holds for fostering greater inclusivity and independence for those who are blind or visually impaired.

2. LITERATURE WORK

In [1],an innovative approach to object detection and tracking that combines Prewitt edge detection with the Kalman filter is introduced. The fundamental objectives of this approach are twofold: first, to represent

the target object effectively, and second, to predict its location accurately, both of which are achieved through the utilization of these algorithms. The implementation of real-time object tracking is demonstrated using a webcam, and our experiments validate the efficiency of our tracking algorithm, showcasing its capability to effectively track moving objects even when they undergo deformation, become occluded, or when multiple objects are present.

In [2], an extensive examination of object detection frameworks founded on deep learning is provided, addressing diverse sub-challenges such as occlusion, clutter, and low-resolution scenarios, often through varying degrees of modifications to the R-CNN framework. The review commences by delving into the fundamentals of generic object detection pipelines, serving as foundational architectures for related tasks. Subsequently, it provides concise overviews of three prevalent tasks: salient object detection, face detection, and pedestrian detection. In conclusion, the paper presents several promising avenues for future exploration aimed at enhancing our comprehension of the object detection domain. This comprehensive review holds significance not only for advancements in neural networks and related learning systems but also offers valuable insights and guidance for future advancements in this field.

In [3], a live object recognition system that serves as a blind aid is introduced. In this system, a Convolutional Neural Network is utilized for the identification of pre-trained objects from the extensive ImageNet dataset. A camera, positioned in accordance with the system's predefined orientation, functions as the input source for a computer system. This system is equipped with the object recognition Neural Network, enabling it to conduct real-time object detection. The results obtained from the network can subsequently be processed to provide information to individuals with visual impairments, either in the form of spoken audio or Braille text.

In [4], a novel approach for vehicle detection in aerial images is presented, centered around an enhanced Faster RCNN model. To address the issue of category imbalance in the training dataset, an oversampling and data stitching augmentation method is introduced, resulting in a balanced dataset. Additionally, the problem of feature loss for small objects caused by pooling operations by enhancing the feature map, which allows to capture more detailed information from the final feature map. Lastly, the method includes a joint training loss function that incorporates center loss for both horizontal and oriented bounding boxes, mitigating the influence of minor inter-class differences on vehicle detection.

In [5], real-time object detection systems YOLOv2 and YOLO9000 are introduced. YOLOv2 represents the state-of-the-art and boasts superior speed compared to other detection systems when applied to various detection datasets. Furthermore, it can be configured to operate at different image sizes, offering a flexible balance between speed and accuracy. On the other hand, YOLO9000 is a real-time framework designed to detect over 9000 object categories by simultaneously optimizing detection and classification. WordTree is used to amalgamate data from diverse sources and employ our joint optimization technique to simultaneously train on ImageNet and COCO. YOLO9000 represents a significant stride toward narrowing the gap in dataset size between detection and classification tasks.

In [6], a novel framework that addresses the challenge of high-quality object detection is introduced. The Cascade R-CNN framework employs a multi-stage cascade structure to progressively filter out false positives and refine object detection results. Through extensive experimentation on diverse datasets, the authors demonstrated significant improvements in accuracy over previous state-of-the-art methods, providing valuable insights into the importance of addressing easy negatives and hard examples at different stages of detection. This work has made a substantial impact on the field of computer vision and object detection, offering a practical solution for real-world applications requiring precise object detection.

In [7], Single Shot MultiBox Detector (SSD), a real-time object detection framework that excels in both speed and accuracy is introduced. The SSD method combines the advantages of deep convolutional neural networks with a set of carefully designed default bounding boxes to predict object categories and locations in a single forward pass. It extensively evaluates the SSD model across various datasets, demonstrating its remarkable efficiency and robustness, particularly for small and overlapping objects. Furthermore, it introduces a multi-scale approach for object detection, allowing the model to handle objects of different sizes and aspect ratios efficiently. This pioneering research has significantly impacted the field of computer vision, serving as a foundation for real-time and accurate object detection, and has inspired subsequent developments in the domain.

In [8], a novel approach that combines multi-region object proposals and semantic segmentation to enhance the accuracy and precision of object detection in introduced. The proposed model integrates convolutional neural networks (CNNs) with region-based object proposal methods, addressing both localization accuracy and class-specific segmentation. It extensively evaluates their approach on benchmark datasets, demonstrating its effectiveness in improving detection results and the capacity to segment object regions accurately. The work has

had a significant impact in advancing the field of computer vision by emphasizing the importance of incorporating semantic segmentation information into object detection pipelines, contributing to the development of more accurate and semantically rich object detection models.

In [9], the authors introduce a novel approach that combines deep convolutional neural networks (CNNs) with a region proposal network, creating a hierarchical model capable of achieving state-of-the-art results in object detection. Their method, known as R-CNN, significantly improves accuracy by generating a rich hierarchy of features and employing selective search for region proposals. Furthermore, the paper addresses the problem of semantic segmentation by treating it as a multi-label classification task and shows how this approach can be integrated into their framework. The research has had a profound impact on the computer vision community, ushering in the era of deep learning for object detection and laying the foundation for numerous subsequent developments in the field, influencing both research and practical applications.

In [10],the authors address the challenge of detecting objects at various scales within images, a fundamental aspect of object detection. They introduce Feature Pyramid Networks (FPN), a novel architecture that enhances the capability of deep convolutional neural networks to efficiently handle objects of different sizes. FPN addresses the issue of information loss due to down-sampling in traditional networks and achieves impressive results by enabling multi-scale feature extraction. By generating feature pyramids and integrating them into object detection pipelines, FPN provides improved object localization and semantic understanding. The research thoroughly evaluates the FPN model across several benchmark datasets, demonstrating its efficacy in improving the accuracy and robustness of object detection systems. This work has had a profound influence on the computer vision community, establishing FPN as a fundamental concept in modern object detection frameworks and setting a standard for designing more effective and scalable deep learning architectures.

3. SYSTEM ARCHITECTURE

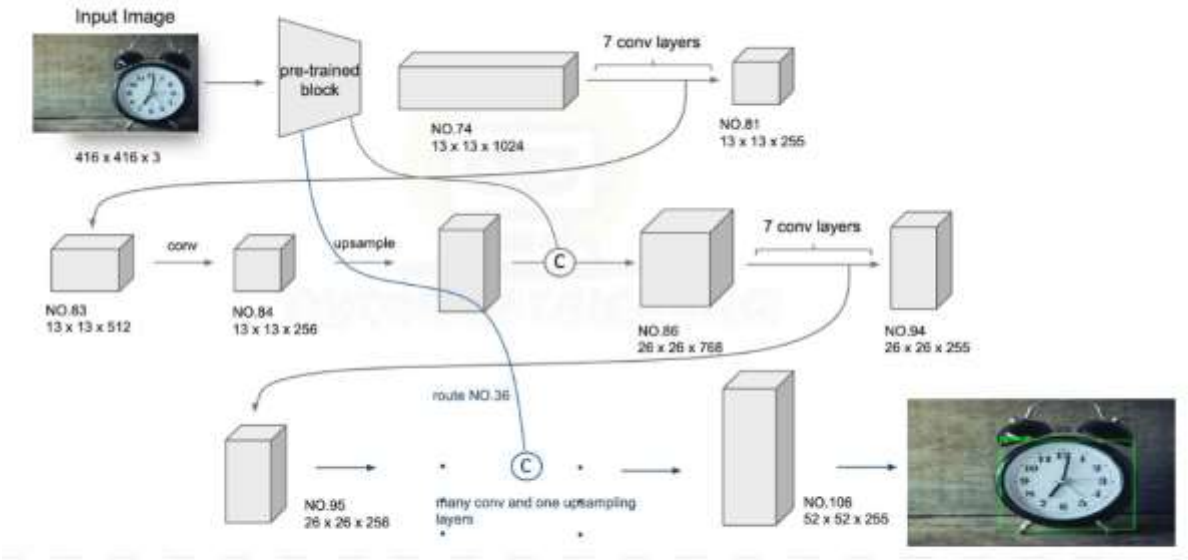


Fig-1: Architecture of yolov3

YOLOv3's architecture can be described as follows:

Input Image

The architecture begins with an input image of a fixed size, which is typically divided into a grid of cells.

Backbone Network

YOLOv3 uses a deep convolutional neural network (CNN) as its backbone to extract features from the input image. The backbone network often consists of variants of Darknet-53, a lightweight neural network architecture.

Detection at Multiple Scales

YOLOv3 performs object detection at three different scales: small, medium, and large. Each scale is associated with a set of anchor boxes that are pre-defined in terms of size and aspect ratio. The detection process at each scale produces bounding box predictions and class probabilities.

Detection Head

At each scale, the detection head generates bounding box coordinates (x, y, width, height) for multiple anchor boxes.It also predicts the class probabilities for objects within each bounding box. The class probabilities are typically estimated using a softmax activation function.

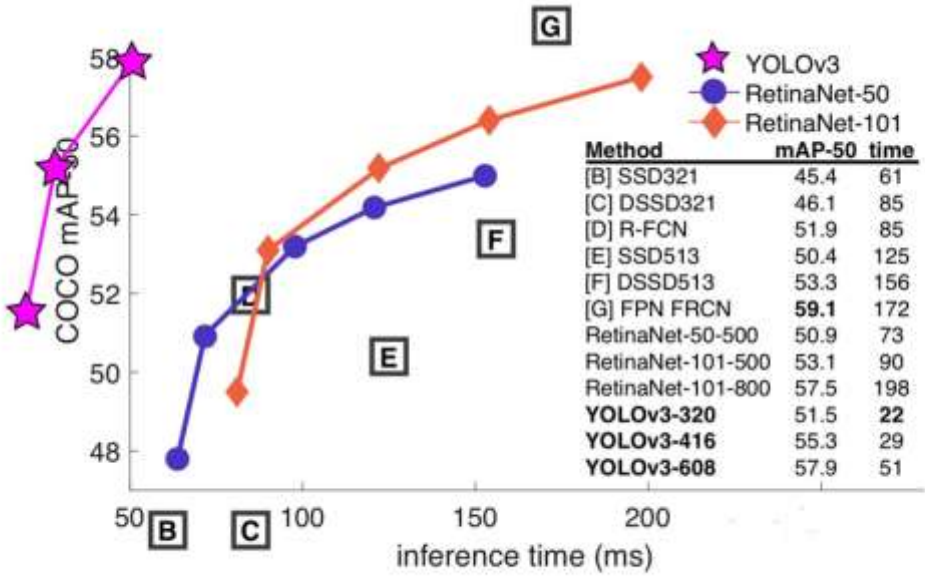
Non-Maximum Suppression (NMS)

After obtaining object predictions at multiple scales, YOLOv3 applies non-maximum suppression to eliminate duplicate and low-confidence detections. NMS ensures that only the most confident and non-overlapping bounding boxes for each object are retained.

Final Object Detection

The results from all scales are combined to create the final set of object detections, which includes the object's class label, confidence score, and bounding box coordinates.

4. RESULTS AND PERFORMANCE ANALYSIS



Method	mAP-50	time
[B] SSD321	45.4	61
[C] DSSD321	46.1	85
[D] R-FCN	51.9	85
[E] SSD513	50.4	125
[F] DSSD513	53.3	156
[G] FPN FRCN	**59.1**	172
RetinaNet-50-500	50.9	73
RetinaNet-101-500	53.1	90
RetinaNet-101-800	57.5	198
YOLOv3-320	51.5	**22**
YOLOv3-416	55.3	29
YOLOv3-608	57.9	51

Fig-2: Comparison of yolov3 with other algorithms

The table shown above represents the comparative performance evaluation of various object detection models on the COCO dataset. The COCO dataset is a large-scale object detection, segmentation, and captioning dataset with 80 object categories.

The first column of the table lists the object detection models being evaluated, including YOLOv3, RetinaNet, Faster R-CNN, Mask R-CNN, and SSD. The second column lists the backbone architecture used for feature extraction in each model. The third column represents the training dataset used to train each model, including

COCO, VOC, and others. The fourth column lists the input image size used during training and inference. The fifth column lists the average precision (AP) obtained by each model on the COCO validation set.

5. CONCLUSION AND FUTURE ENHANCEMENT

Implementing YOLOv3-based object detection for visually impaired individuals offers substantial benefits, enabling real-time and accurate object recognition. YOLOv3's speed and versatility make it a valuable tool for enhancing the mobility and safety of visually impaired individuals. It provides immediate feedback, aids in identifying common obstacles, and empowers users to navigate their surroundings more independently.

To further enhance YOLOv3 for this context, future efforts should focus on refining object classification, improving localization accuracy, integrating with navigation systems, and creating user-friendly interfaces with auditory feedback. Customization and regular updates to accommodate evolving needs are crucial for making the technology more user-centric and effective in assisting visually impaired individuals in their daily lives.

6. REFERENCES

[1] S. Cherian and C. Singh, "Real Time Implementation of Object Tracking Through webcam" in International Journal of Research in Engineering and Technology 128–132, Oxford:Clarendon, vol. 2, pp. 68-73, 2014.

[2] Z. Zhao, Q. Zheng, P. Xu, S. T and X. Wu, "Object detection with deep learning: A review", IEEE transactions on neural networks and learning systems, vol. 30, no. 11, pp. 3212-3232, 2019.

[3] K. Potdar, C. D. Pai and S. Akolkar, "A convolutional neural network based live object recognition system as blind aid", arXiv preprint arXiv:1811.10399, 2018.

[4] Mo Nan and Li Yan, "Improved Faster RCNN Based on Feature Amplification and Oversampling Data Augmentation for Oriented Vehicle Detection in Aerial Images", Remote Sensing 12, no. 16, pp. 2558, 2020.

[5] Redmon Joseph and Ali Farhadi, "YOLO9000: better faster stronger", Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263-7271, 2017.

[6] Cai Zhaowei and Nuno Vasconcelos, "Cascade r-cnn: Delving into high quality object detection", Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6154-6162, 2018.

[7] Liu Wei, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, et al., "Ssd: Single shot multibox detector", European conference on computer vision, pp. 21-37, 2016.

[8] Gidaris Spyros and Nikos Komodakis, "Object detection via a multi-region and semantic segmentation-aware cnn model", Proceedings of the IEEE international conference on computer vision, pp. 1134-1142, 2015.

[9] Ross Girshick, Donahue Jeff, Darrell Trevor and Malik Jitendra, "Rich feature hierarchies for accurate object detection and semantic segmentation", Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587, 2014.

[10] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan and Serge Belongie, "Feature pyramid networks for object detection", CVPR, vol. 1, no. 2, pp. 4, 2017.