International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

# Object Detection with Audio Feedback

**K. Omsainath[1], Mrs. P. Poornima [2]**
**MCA Student[1], Assistant Professor[2]**
Mother Theresa Institute of Computer Applications, Palamaner , S.V University, Titupathi, Andhra Pradesh,, India

**Article Info**
**Publication Issue :**
Volume 8, Issue 6
November-December-2022
**Page Number :** 12-18
**Article History**
Accepted: 01 Nov 2022
Published: 03 Nov 2022

**ABSTRACT**

Object recognition is one of the challenging application of computer vision, which has been widely applied in many areas for e.g. autonomous cars, Robotics, Security tracking, Guiding Visually Impaired Peoples etc. With the rapid development of deep learning many algorithms were improving the relationship between video analysis and image understanding. All these algorithms work differently with their network architecture but with the same aim of detecting multiple objects within complex image. Absence of vision impairment restraint the movement of the person in an unfamiliar place and hence it is very essential to take help from our technologies and trained them to guide blind peoples whenever they need.

Keywords: Tensor flow, Yolo_v3, Web Speech API, Deep Learning.

## I. INTRODUCTION

Humans almost by birth are trained by their parents to categorize between various objects as children self is one object. Human Visual System is very accurate and precise that can handle multi-tasks even with less conscious mind. When there is large data then we need more accurate system to correctly recognize and localize multiple objects simultaneously. Here machines comes into existence, we can train our computers with the help of better algorithms to detect multiple objects within the image with high accuracy and preciseness. Object Detection is the most challenging application of computer vision as it require complete understanding of images. In other words object tracker tries to find the presence of object within multiple frames and assigns labels to each object. There might be many problems faced by the tracker in terms of complex image, Loss of information and transformation of 3D world into 2 D image. To achieve good accuracy in object detection we should not only focus on classifying objects but also on locating the positions of different objects that may vary image to image. It is very important to develop the most effective real time object tracking algorithm which is a challenging task. Deep learning since 2012 is working in these kinds of problems and has revolutionized the domain of computer vision. This paper aims to test the performance of both the algorithms in different situations in real time using webcam and is made primarily for the visually impaired peoples. Blind peoples have to rely on someone who can guide them or on their physical touch which is sometimes very risky also. Daily navigation of blind peoples in unfamiliar environments could be the frighten task without the

help of some intelligent systems. They key concern behind this contribution is to investigate the possibility of expanding the counts of objects at one go to expand the support given to the visually impaired peoples. Some common limitations of the previous techniques is less accuracy, complexity in scene, lightening etc. To overcome all those challenges two algorithms are analyzed on all possible grounds and from every perspective to achieve good accuracy.

YOLO as the name suggest "You Only Look Once" is the first algorithm after selective search approach that uses single neural network to the image and then divide the image into SXS grids and create the bounding box by assigning the confidence score and class label to each object. Each grid cell is predicting (x, y, w, h) and confidence score for every object. Confidence score measures how accurately the object is present inside the bounding box. The value of all (x, y, w, h) are between 0 and 1. YOLO prediction has a shape of (S, S, BX5+C). Network Architecture of Yolo algorithm with counting of each convolutional layer can be studied from every grid has 20 class conditional probability. Class conditional probability tells the probability of object presence in the cell. The above description covers basic features of YOLO algorithm, for deep understanding refer in developing Object detection using YOLO with voice response we have used tensor flow with SSD _Mobile net model. Single Shot Detector (SSD). Tensor flow is a google open source machine framework which is used to detect real world objects in a frame. Tensor flow is a unique model that is used according to the need of the user. This library of python uses many architectures for e.g. SSD (Single hot Detector), CNN, Faster CNN etc. these play very important role in speed and accuracy of model. In Building our model we have used tensor flow_SSD _Mobile Net model. SSD as the name suggest detect the class and position of the object in same step. Mobile Net is a convolutional feature extractor used to extract high level features of the images. Once the class and position of multiple objects is detected by feeding the dataset in the YOLO architecture using Tensor flow model the text output is converted into speech using "gTTS".

## II. Related works

**Real time implementation of object tracking through webcam:** Real time object detection and tracking is an important task in various computer vision applications. For robust object tracking the factors like object shape variation, partial and full occlusion, scene illumination variation will create significant problems. We introduce object detection and tracking approach that combines Prewitt edge detection and kalman filter. The target object's representation and the location prediction are the two major aspects for object tracking this can be achieved by using these algorithms. Here real time object tracking is developed through webcam. Experiments show that our tracking algorithm can track moving object efficiently under object deformation, occlusion and can track multiple objects.

**Object Detection with Deep Learning: A Review:** Due to object detection's close relationship with video analysis and image understanding, it has attracted much research attention in recent years. Traditional object detection methods are built on handcrafted features and shallow trainable architectures. Their performance easily stagnates by constructing complex ensembles which combine multiple low-level image features with high-level context from object detectors and scene classifiers. With the rapid development in deep learning, more powerful tools, which are able to learn semantic, high-level, deeper features, are introduced to address the problems existing in traditional architectures. These models behave differently in network architecture, training strategy and optimization function, etc. In this paper, we provide a review on deep learning based object detection frameworks. Our review begins with a brief

introduction on the history of deep learning and its representative tool, namely Convolutional Neural Network (CNN). Then we focus on typical generic object detection architectures along with some modifications and useful tricks to improve detection performance further. As distinct specific detection tasks exhibit different characteristics, we also briefly survey several specific tasks, including salient object detection, face detection and pedestrian detection. Experimental analyses are also provided to compare various methods and draw some meaningful conclusions. Finally, several promising directions and tasks are provided to serve as guidelines for future work in both object detection and relevant neural network based learning systems.

**Histograms of oriented gradients for human detection:** We study the question of feature sets for robust visual object recognition; adopting linear SVM based human detection as a test case. After reviewing existing edge and gradient based descriptors, we show experimentally that grids of histograms of oriented gradient (HOG) descriptors significantly outperform existing feature sets for human detection. We study the influence of each stage of the computation on performance, concluding that fine-scale gradients, fine orientation binning, relatively coarse spatial binning, and high-quality local contrast normalization in overlapping descriptor blocks are all important for good results. The new approach gives near-perfect separation on the original MIT pedestrian database, so we introduce a more challenging dataset containing over 1800 annotated human images with a large range of pose variations and backgrounds.

**Region-Based Convolutional Networks for Accurate Object Detection and Segmentation:** Object detection performance, as measured on the canonical PASCAL VOC Challenge datasets, plateaued in the final years of the competition. The best-performing methods were complex ensemble systems that typically combined multiple low-level image features with high-level context. In this paper, we propose a simple and scalable detection algorithm that improves mean average precision (MAP) by more than 50 percent relative to the previous best result on VOC 2012-achieving a MAP of 62.4 percent. Our approach combines two ideas: (1) one can apply high-capacity convolutional networks (CNNs) to bottom-up region proposals in order to localize and segment objects and (2) when labeled training data are scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, boosts performance significantly. Since we combine region proposals with CNNs, we call the resulting model an R-CNN or Region-based Convolutional Network.

**YOLO9000: better, faster, stronger:**

We introduce YOLO9000, a state-of-the-art, real-time object detection system that can detect over 9000 object categories. First we propose various improvements to the YOLO detection method, both novel and drawn from prior work. The improved model, YOLOv2, is state-of-the-art on standard detection tasks like PASCAL VOC and COCO. Using a novel, multi-scale training method the same YOLOv2 model can run at varying sizes, offering an easy tradeoff between speed and accuracy. At 67 FPS, YOLOv2 gets 76.8 mAP on VOC 2007. At 40 FPS, YOLOv2 gets 78.6 mAP, outperforming state-of-the-art methods like Faster RCNN with ResNet and SSD while still running significantly faster. Finally we propose a method to jointly train on object detection and classification. Using this method we train YOLO9000 simultaneously on the COCO detection dataset and the ImageNet classification dataset. Our joint training allows YOLO9000 to predict detections for object classes that dont have labelled detection data. We validate our approach on the ImageNet detection task. YOLO9000 gets 19.7 mAP on the ImageNet detection validation set despite only having detection data for 44 of the 200 classes. On the 156 classes not in COCO, YOLO9000 gets 16.0 mAP. YOLO9000 predicts detections for more than 9000 different object categories, all in real-time.

**Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks:** State-of-the-art object detection networks depend on region proposal algorithms to hypothesize object locations. Advances like SPPnet and Fast R-CNN have reduced the running time of these detection networks, exposing region proposal computation as a bottleneck. In this work, we introduce a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. An RPN is a fully convolutional network that simultaneously predicts object bounds and objectness scores at each position. The RPN is trained end-to-end to generate high-quality region proposals, which are used by Fast R-CNN for detection. We further merge RPN and Fast R-CNN into a single network by sharing their convolutional features—using the recently popular terminology of neural networks with "attention" mechanisms, the RPN component tells the unified network where to look. For the very deep VGG-16 model, our detection system has a frame rate of 5fps (including all steps) on a GPU, while achieving state-of-the-art object detection accuracy on PASCAL VOC 2007, 2012, and MS COCO datasets with only 300 proposals per image. In ILSVRC and COCO 2015 competitions, Faster R-CNN and RPN are the foundations of the 1st-place winning entries in several tracks. Code has been made publicly available.

## III. Methodology

**Proposed system:**

We propose a system that will detect every possible day to day multiple objects on the other hand prompt a voice to alert person about the near as well as farthest objects around them. To get audio we will use web speech API to produce speech.

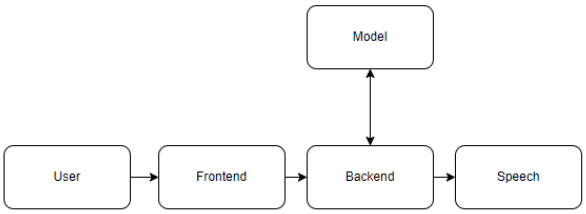



**Figure 1: Block diagram**

## IV. Implementation

The project has implemented by using below listed algorithm.

**Conventional Neural Network (CNN):**

A convolutional neural network consists of an input layer, hidden layers and an output layer. In any feed-forward neural network, any middle layers are called hidden because their inputs and outputs are masked by the activation function and final convolution. In a convolutional neural network, the hidden layers include layers that perform convolutions. Typically this includes a layer that performs a dot product of the convolution kernel with the layer's input matrix. This product is usually the Frobenius inner product, and its activation function is commonly ReLU. As the convolution kernel slides along the input matrix for the layer, the convolution operation generates a feature map, which in turn contributes to the input of the next layer. This is followed by other layers such as pooling layers, fully connected layers, and normalization layers.

**Convolutional layers**

In a CNN, the input is a tensor with a shape: (number of inputs) x (input height) x (input width) x (input channels). After passing through a convolutional layer, the image becomes abstracted to a feature map, also called an activation map, with shape: (number of inputs) x (feature map height) x (feature map width) x (feature map channels).

Convolutional layers convolve the input and pass its result to the next layer. This is similar to the response of a neuron in the visual cortex to a specific stimulus. Each convolutional neuron processes data only for its receptive field. Although fully connected feed forward neural networks can be used to learn features and classify data, this architecture is generally impractical for larger inputs such as high resolution images. It would require a very high number of neurons, even in a shallow architecture, due to the large input size of images, where each pixel is a relevant input feature. For instance, a fully connected layer for a (small) image of size 100 x 100 has 10,000 weights for each neuron in the second layer. Instead, convolution reduces the number of free parameters, allowing the network to be deeper. For example, regardless of image size, using a 5 x 5 tiling region, each with the same shared weights, requires only 25 learnable parameters. Using regularized weights over fewer parameters avoids the vanishing gradients and exploding gradients problems seen during back propagation in traditional neural networks. Furthermore, convolutional neural networks are ideal for data with a grid-like topology (such as images) as spatial relations between separate features are taken into account during convolution and/or pooling.

**Pooling layers**

Convolutional networks may include local and/or global pooling layers along with traditional convolutional layers. Pooling layers reduce the dimensions of data by combining the outputs of neuron clusters at one layer into a single neuron in the next layer. Local pooling combines small clusters, tiling sizes such as 2 x 2 are commonly used. Global pooling acts on all the neurons of the feature map. There are two common types of pooling in popular use: max and average. Max pooling uses the maximum value of each local cluster of neurons in the feature map, while average pooling takes the average value.

**Fully connected layers**

Fully connected layers connect every neuron in one layer to every neuron in another layer. It is the same as a traditional multi-layer perceptron neural network (MLP). The flattened matrix goes through a fully connected layer to classify the images.

**Receptive field**

In neural networks, each neuron receives input from some number of locations in the previous layer. In a convolutional layer, each neuron receives input from only a restricted area of the previous layer called the neuron's receptive field. Typically the area is a square (e.g. 5 by 5 neurons). Whereas, in a fully connected layer, the receptive field is the entire previous layer. Thus, in each convolutional layer, each neuron takes input from a larger area in the input than previous layers. This is due to applying the convolution over and over, which takes into account the value of a pixel, as well as its surrounding pixels. When using dilated layers, the number of pixels in the receptive field remains constant, but the field is more sparsely populated as its dimensions grow when combining the effect of several layers.

**Weights:**

Each neuron in a neural network computes an output value by applying a specific function to the input values received from the receptive field in the previous layer. The function that is applied to the input values is determined by a vector of weights and a bias (typically real numbers). Learning consists of iteratively adjusting these biases and weights.

The vector of weights and the bias are called filters and represent particular features of the input (e.g., a particular shape). A distinguishing feature of CNNs is that many neurons can share the same filter. This reduces the memory footprint because a single bias and a single vector of weights are used across all receptive fields that share that filter, as opposed to

each receptive field having its own bias and vector weighting.

**YOLO:**

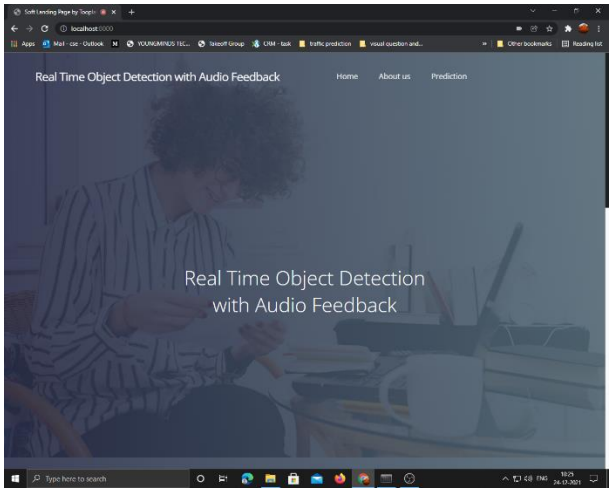
Yolo is a part of object detection, Object detection is a computer technology related to computer vision and image processing that deals with detecting instances of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and videos. Well-researched domains of object detection include face detection and pedestrian detection. Object detection has applications in many areas of computer vision, including image retrieval and video surveillance.

Every object class has its own special features that helps in classifying the class – for example all circles are round. Object class detection uses these special features. For example, when looking for circles, objects that are at a particular distance from a point (i.e., the center) are sought. Similarly, when looking for squares, objects that are perpendicular at corners and have equal side lengths are needed. A similar approach is used for face identification where eyes, nose, and lips can be found and features like skin color and distance between eyes can be found.

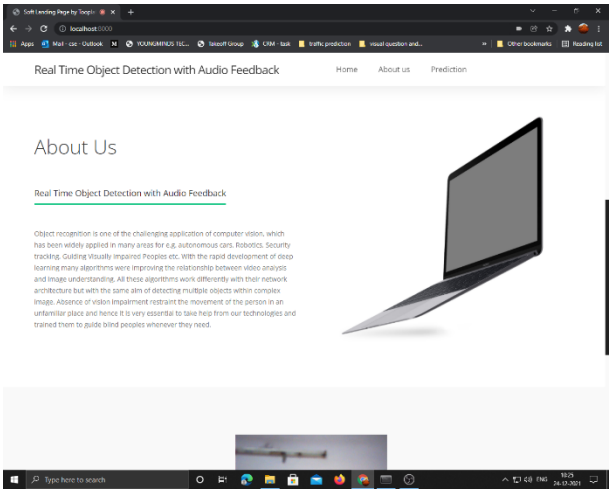
## V. Results and Discussion

The following screenshots are depicted the flow and working process of project.

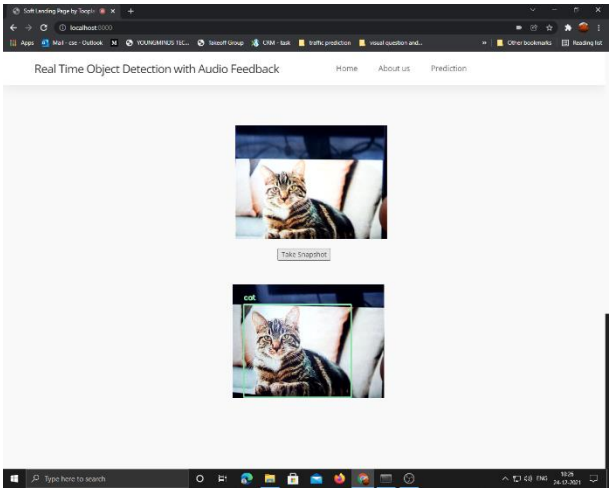
**Home Page:** This is the home page of real time object detection with audio feedback.




**About Page:** This page describes the concept behind this project




**Prediction:** This page displays predicted output.




## VI. Conclusion

In this project we have developed a user friendly application called the detection of objects via audio feedback. We provide a system that will identify all conceivable daily numerous things and then urge a voice to warn a person about the closest and farthest objects nearby. We'll utilize the online speech API to generate speech to obtain audio.

## VII. References

[1]. S. Cherian, & C. Singh, "Real Time Implementation of Object Tracking Through

webcam," Internation Journal of Research in Engineering and Technology, 128-132, (2014).

[2]. Z. Zhao, Q. Zheng, P.Xu, S. T, & X. Wu, "Object detection with deep learning: A review," IEEE transactions on neural networks and learning systems, 30(11), 3212-3232, (2019).

[3]. N. Dalal, & B. Triggs, "Histograms of oriented gradients for human detection," In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 886-893). IEEE, (2005, June).

[4]. R. Girshick., J. Donahue, T. Darrell, & J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," IEEE transactions on pattern analysis and machine intelligence, 38(1), 142-158, (2015).

[5]. X. Wang, A. Shrivastava, & A. Gupta, "A-fast-rcnn: Hard positive generation via adversary for object detection," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2606- 2615), (2017).

[6]. S. Ren, K. H, R. Girshick, & J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," In Advances in neural information processing systems (pp. 91-99), (2015).

[7]. J. Redmon, S. Divvala, R. Girshick, & A. Farhadi, "You only look once: Unified, real-time object detection," In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788), (2016).

[8]. J. Redmon, & A. Farhadi, "YOLO9000: better, faster, stronger," In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271) (2017).

[9]. J. Redmon & A. Farhadi, "Yolov3: An incremental improvement," ArXiv preprint arXiv: 1804.02767, (2018).

[10]. R. Bharti, K. Bhadane, P. Bhadane, & A. Gadhe, "Object Detection and Recognition for Blind Assistance," International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 06, (2019).

[11]. T. Lin, Y. Maire, M. Belongie, S. Hays, J. Perona, P. Ramanan, D., & C.L. Zitnick, "Microsoft coco: Common objects in context," In European conference on computer vision (pp. 740-755). Springer, Cham, (2014, September).

[12]. J. Du, "Understanding of Object Detection Based on CNN Family and YOLO" in Journal of Physics: Conference Series, IOP Publishin, g, vol. 1004, no. 1, pp. 012029, April 2018.

[13]. S. Geethapriya, N. Duraimurugan and S.P. Chokkalingam, "Real-Time Object Detection with Yolo", *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 8, no. 3S, 2019.

[14]. A. Arora, A. Grover, R. Chugh and S.S. Reka, "Real time multi object detection for blind using single shot multibox detector", *Wireless Personal Communications*, vol. 107, no. 1, pp. 651-661, 2019.

[15]. S. Kurleka, "Reading Device for Blind People using Python OCR", *International Journal of Science and Engineering Applications*, vol. 9, no. 04, pp. 49-52, 2020, ISSN 2319-7560.

[16]. J. Li, J. Gu, Z. Huang and J. Wen, "Application Research of Improved YOLO V3 Algorithm in PCB Electronic Component Detection", *Applied Sciences*, vol. 9, no. 18, pp. 3750, 2019.

[17]. A. P. Jana and A. Biswas, "YOLO based Detection and Classification of Objects in video records", *2018 3rd IEEE International Conference on Recent Trends in Electronics Information & Communication Technology (RTEICT)*, pp. 2448-2452, 2018, May.

[18]. G. Peng, "Performance and Accuracy Analysis in Object Detection", 2019.

**Cite this article as :**

K. Omsainath, Mrs. P. Poornima, "Object Detection with Audio Feedback", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 6, pp. 12-19, November-December 2022.

Journal URL : https://ijsrcseit.com/CSEIT228559