International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 12 Issue 3, March-2023, Impact Factor: 7.957

# Real Time Object Detection with Audio Feedback Using Yolo_V3

M. Ujjwala[1], K. Venkatabharath[2], M. Sudeep[3], C. Venkata Ramana[4], B. Vijaya Durga[5], Dr. P. Sudhakara Reddy[6]

[1,2,3,4,5]Department of Computer Science and Engineering, Annamacharya Institute of Technology and Sciences(Autonomous), Rajampet, Andhra Pradesh, India-516126
[6]Professor in Department of CSE, Annamacharya Institute of Technology and Sciences (Autonomous), Rajampet, Andhra Pradesh, India-516126

**ABSTRACT**

**With the advent of Information Technology, the object recognition is one of the challenging applications. In particular, object recognition is widely used computer vision applications such as autonomous cars, robotics, security tracking, guiding visually impaired people etc. In this, mainly we study image understanding of computer vision and analysis plays a key role in present application scenario. In this project work, we proposed a simple task to extract information about the object i.e.,a sample image and its detection using yolo_v3 algorithm. In continuation, we are taken trained data image samples(nearly 300) to detect multiple objects. The application of project proposed to guide blind people whenever they need i.e., in moving streets and busy areas, etc. Here, a voice signal is used to alert person about the near as well as farthest objects around them. An image to text to speech to audio conversion process was done by using gTTs (Google Text to Speech) conversion tool. A python programme is used to detect object and proposed to implement by using Yolo_v3 tool. The verification of accuracy and performance parameters was noted and tested successfully. Testing was performed by using MS-COCO Dataset samples.**

**Keywords: Deep Learning, gTTS, SSD, Tensor flow, Yolo_v3.**

## INTRODUCTION

With the recent rapid development of information technology (IT), a lot of research has been carried out to solve inconveniences in everyday life, and as a result, various conveniences for people have been provided. Nevertheless, the visually challenged still face a lot of challenges. One of the identified inconvenience is that a blind person feels in everyday life include finding information about objects and indoor mobility problems. They have difficulty recognizing simple objects, and it is not easy to distinguish objects that have similar forms. Blind people find it difficult to identify various objects compared to other human beings.

Now that machines have been created, we may train our computers to recognise several items in an image with accuracy and precision by using YOLO version 3platform other than like R-CNN, HOG etc. The most difficult use of computer vision is object detection since it necessitates a thorough comprehension of images. In other words, an object tracker looks for the presence of objects across several frames and identifies them individually. The tracker may experience numerous issues with complicated images, information loss, and the conversion of 3D environments into 2D images. We need not only concentrate on identifying items, but also on locating the positions of many objects that may vary from image to image if we want to attain good accuracy in object detection.

It is crucial to create the best real-time object tracking algorithm, which is a difficult challenge. The efforts are made to design this project particularly for those who are visually impaired and intends to test programme's functionality in a variety of settings using a webcam in real time. Without the aid of some intelligent devices, blind people's daily navigation of unfamiliar situations like navigating through the new environments that they are not familiar with, could be a terrifying task. The existing yolo_v3 algorithm is examined from every angle and on every conceivable basis to overcome the obstacles and attain high accuracy and help to guide the visually impaired people properly in an unfamiliar environment.

## DATA SET

Common Objects In Context (COCO) dataset is used to train the model. The model here is the You Only Look Once_Version3 (YOLO_V3) algorithm that runs through a variation of an extremely complex Convolutional Neural

Network architecture called the Darknet. The two key qualities we consider whenever we work on an object detection algorithm are detection and localisation. Detection refers to the ability of the algorithm to accurately identify the presence and type of objects in an image or video and also distinguishing it from other objects or background elements. Whether an item belongs to a particular class or not must be stated when detecting it. Localization, on the other hand, refers to the ability of the algorithm to accurately locate the object within the image or video, as the location of an object can vary from image to image, localization refers to the bounding box around every object. This allows for further analysis or processing of the object, such as tracking its movement over time or extracting features for use in other applications.

The use of challenging datasets can be an effective way to set benchmarks for measuring the performance of programs. Challenging datasets can provide a range of difficult scenarios that an algorithm may encounter in real-world applications, such as occlusions, variations in lighting or background, and changes in object appearance or position.

Our system was trained to detect the following objects - ['traffic light', 'oven', 'keyboard', 'kite', 'sandwich', 'dog', 'potted plant', 'toilet','toaster', 'cell phone', 'clock', 'bicycle', 'bicycle', 'cell phone', 'toothbrush', 'kite', 'broccoli', 'fire hydrant', 'umbrella', 'umbrella', 'parking meter', 'keyboard', 'bench', 'sink', 'sheep', 'giraffe', 'keyboard', 'handbag', 'pizza', 'book', 'fork', 'book', 'clock', 'bench', 'hair drier', 'dog', 'toaster', 'horse','giraffe', 'car', 'umbrella', 'truck', 'wine glass', 'handbag', 'bus', 'banana', 'bench', 'donut', 'clock', 'knife', 'sheep', 'skateboard', 'bus', 'toothbrush', 'traffic light', 'cake', 'airplane', 'person', 'chair', 'suitcase', 'cell phone', 'dog', 'laptop', 'apple', 'motorcycle', 'scissors', 'oven', 'banana', 'cake', 'vase', 'bowl', 'parking meter', 'zebra', 'cat', 'cell phone', 'tie', 'horse', 'cow', 'potted plant', 'hair drier', 'mouse', 'sandwich', 'broccoli', 'oven', 'surfboard', 'carrot', 'bowl', 'bottle', 'suitcase', 'zebra', 'scissors', 'book', 'traffic light', 'teddy bear', 'clock', 'wine glass', 'kite', 'scissors', 'sandwich', 'bear', 'handbag', 'sports ball', 'orange', 'sandwich', 'spoon', 'surfboard', 'bench', 'bottle', 'dog', 'mouse', 'cell phone', 'toaster', 'banana', 'oven', 'dining table', 'fork', 'apple', 'baseball bat', 'frisbee', 'book', 'tie', 'scissors', 'mouse', 'car', 'skateboard', 'sheep', 'knife', 'wine glass', 'airplane', 'skateboard', 'wine glass', 'elephant', 'cat', 'cake', 'refrigerator', 'handbag', 'potted plant', 'toilet', 'laptop', 'toaster', 'cat', 'bottle', 'backpack', 'bear', 'traffic light', 'boat', 'sheep', 'sandwich', 'cup', 'tv', 'horse', 'bowl', 'hot dog', 'baseball bat', 'sheep', 'bear', 'cup', 'teddy bear', 'stop sign', 'baseball bat', 'laptop', 'apple', 'kite', 'wine glass', 'snowboard', 'dining table', 'motorcycle', 'suitcase', 'pizza', 'tennis racket', 'motorcycle', 'dog', 'laptop', 'frisbee', 'couch', 'sports ball', 'bicycle', 'tie', 'bear', 'frisbee', 'keyboard', 'umbrella', 'motorcycle', 'tv', 'clock', 'chair', 'refrigerator', 'dining table', 'keyboard', 'motorcycle', 'teddy bear', 'cup', 'couch', 'skis', 'bench', 'sink', 'dining table', 'horse', 'clock', 'bear', 'carrot'].

## METHODOLOGY

This section presents our proposed methodology for detecting the real-time objects from the input images by using convolutional neural network. Convolutional neural networks (CNNs) are a class of deep neural networks used most frequently for the analysis of visual vision in deep learning. The previous algorithms such as CNN, faster CNN, faster RCNN, SSD are only suitable for highly powerful computing machines and they require a large amount of time to train. The proposed scheme uses YOLO v_3 algorithm for higher detection precision with real-time speed. The key components of the YOLO v_3 algorithm is presented in this section. With the aid of a diagram, the framework of the algorithm for identifying multiple objects with audible feedback is demonstrated, providing a clear grasp of the real flow of the process.

**Real Time Object Detection with Audio Feedback using Yolo_v3 Algorithm**

When we need real-time detection, You Only Look Once is one of the fastest object detection algorithms. The YOLO algorithm just released its third version, known as YOLO V3, which is more faster and more accurate than Single Shot Detector (SSD). YOLO V3 originally had 53 layers of dark net, however 53 layers were added for detection, giving us a total of 106 layers. In the enhanced architecture, YOLO V3 algorithm's ability to make detections at three different scales, three different sizes, and three different locations throughout the network is its most intriguing feature.

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. The YOLO_V3 architecture classifies a number of items after feeding the input image, and each object is given a class label. As seen in "Fig. 1" to get the audio feedback gTTs(Google Text to Speech), python library used to convert statements to audio speech. To play the audio pygame python module is used.

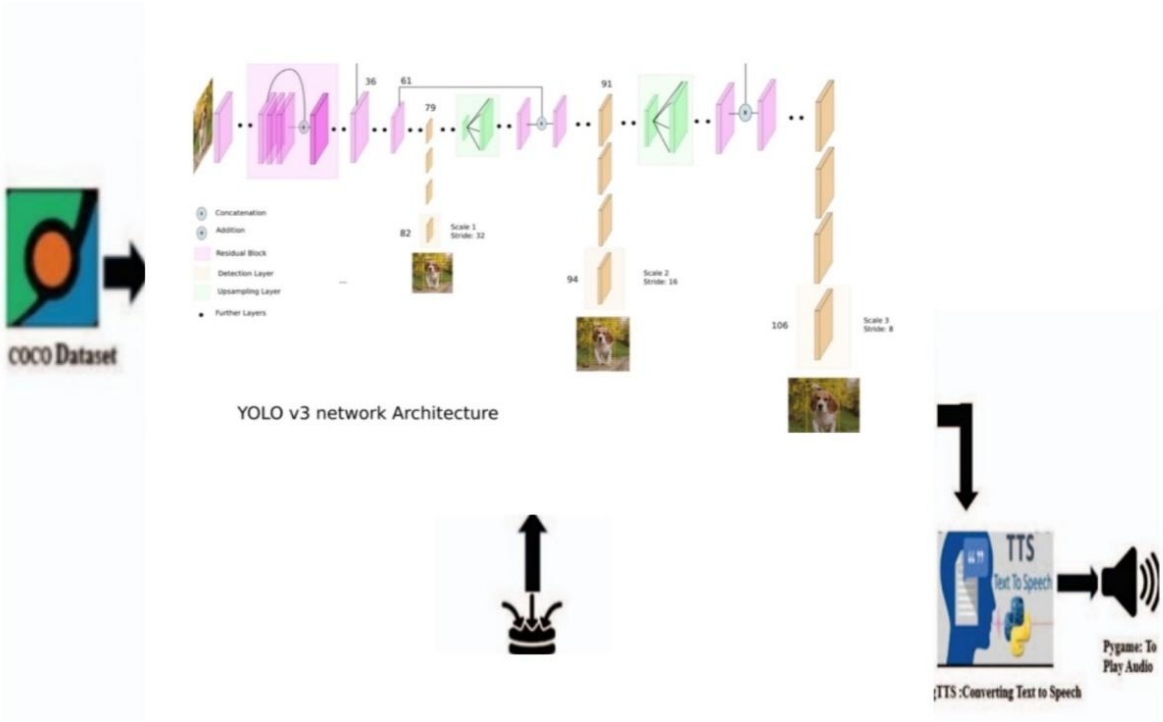


**Fig 1.Workflow of YOLO_V3 with Audio Feedback**

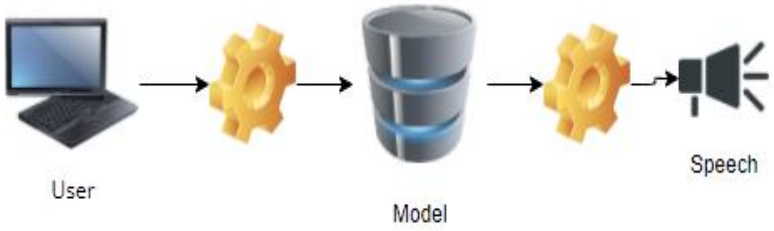


**Fig 2.ARCHITECTURE**

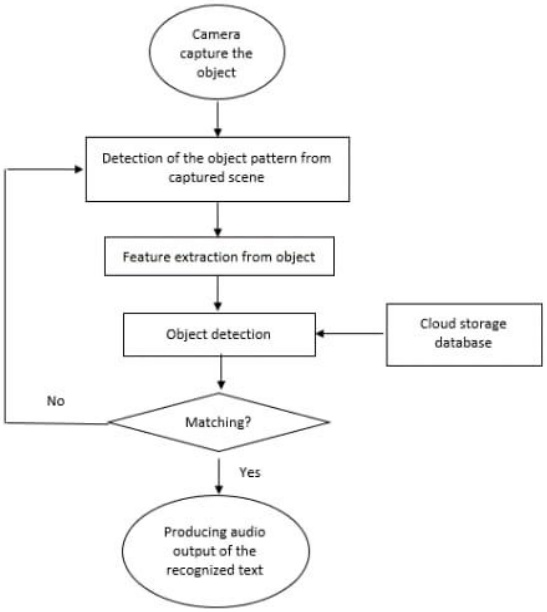


**Fig 3.Flow chart of the object detection**

## EXPRIMENTAL RESULTS AND ANALYSIS

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 instance, when looking for circles, one seeks things that are a certain distance away from a point (i.e., the centre). Similar to searching for squares, finding items with equal side lengths and perpendicular corners is necessary. Similar methods are employed to identify face using traits including skin tone and the distance between eyes as well as nose, lips.

Similarly, in this work the system to trained to detect around 200 objects. For the implementation, with the pycharm IDE in the windows operating system, the Django web framework is used. It is an advanced Python web framework that enables quick creation of safe and dependable websites. From Django, the packages required are imported and we set a path to get the object details mentioning the file name and the model used is "yolo_v3" where around 200 classes are available using COCO's pretrained weights. By taking a snapshot of the image given as the input is read and processed using the yolo_v3 algorithm. After feeding the input image in the yolo_v3 architecture, the input is classified and bounded with the class labels and bounding boxes to map with the data set elements. If the objects in the input given by the user are mapped to the data set, it predicts the output and provides a voice note regarding the object detected in the image. In an information system, the raw data that is processed to produce output is referred to as input. As a result, the system output's quality depends on the input's quality.

The resulting output is sent to get the audio feedback gTTS(Google Text to Speech), python library is used to convert statements into audio speech. Pygame python module is used to play the audio. To compile the code, the command – python manage.py runserver is entered and the webpage appears as shown in Fig. 4.

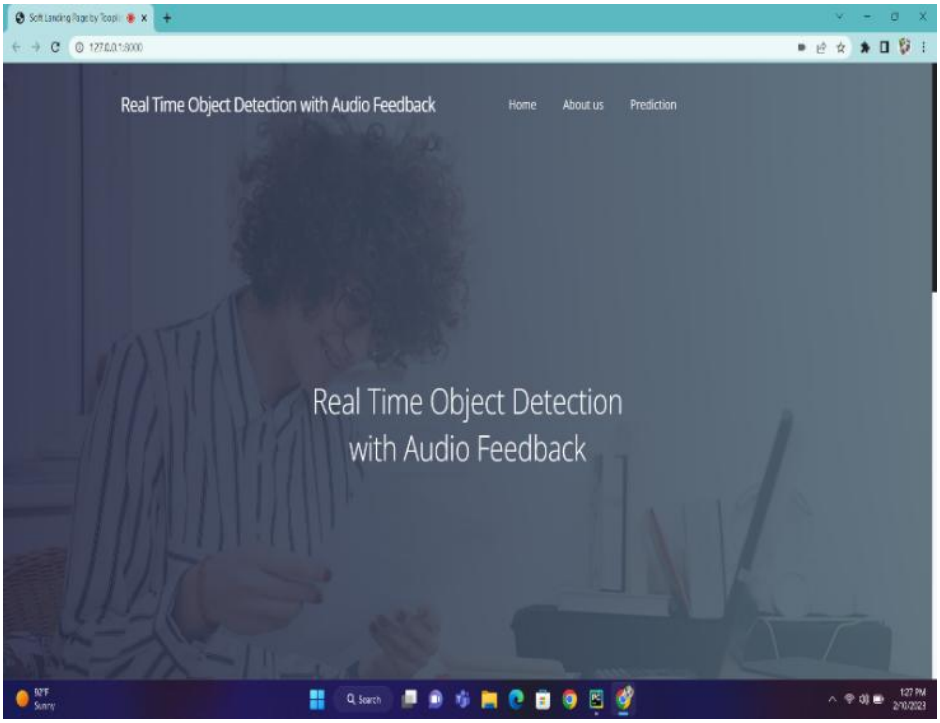


**Fig 4.The webpage displays where the input is given by the user**

All the experiments below were done in real time using webcam in various situations i.e., Single object objection, multiple object detection.

## SINGLE OBJECT OUTPUT

YOLO_V3 provides good accuracy in the single object test. It provides a voice prompt after detecting the object. The single object detection is illustrated in the following examples. In the Fig. 5,a snapshot of single object as person is taken and the output is the audio feedback of the object detected as 'person'. In the Fig. 6, In a classroom of AITS college, the single object bench is given as input, and the output generated is the audio as 'Bench'. In another example, Fig. 7,a snapshot of the object laptop is taken, as the object is detected it alerts with a voice message as 'Laptop'.

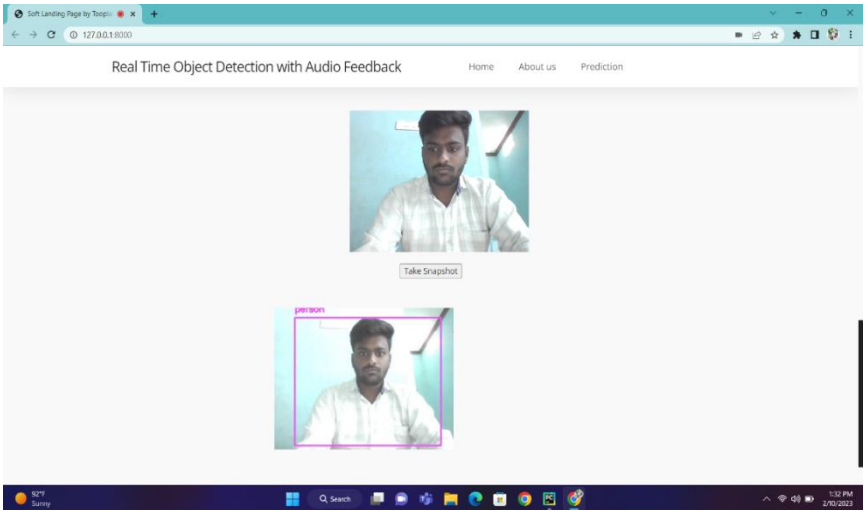


**Fig 5.A snapshot of the image is taken and the output is the voice prompt as person**

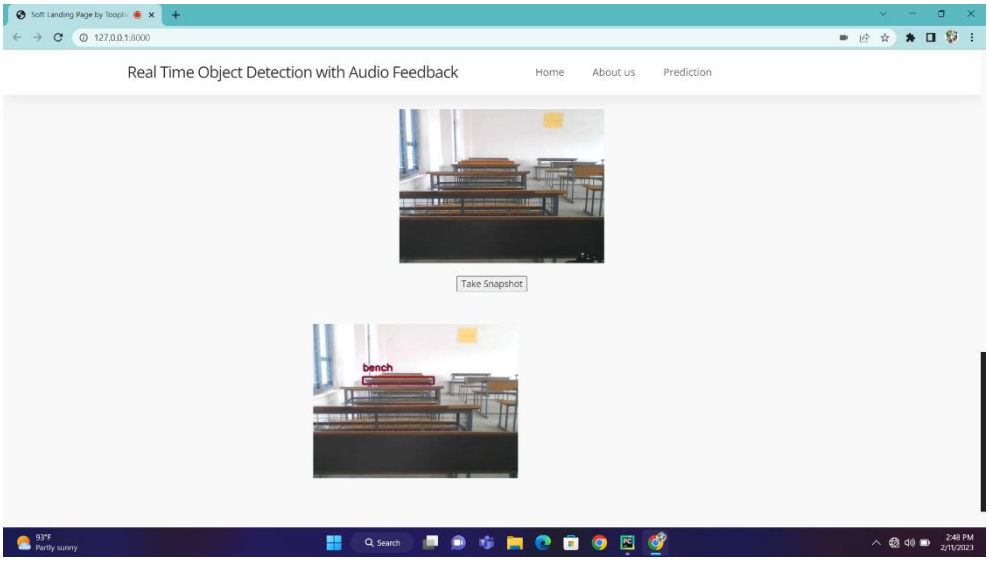


**Fig 6.In a classroom of AITS college, a snapshot is taken and it prompts a voice note as bench as it is detected.**

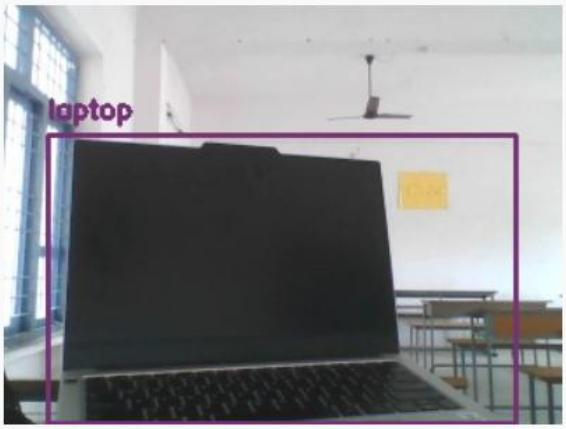


**Fig7.An input is given as a single object laptop and the output after detection is the voice alert as laptop**

## MULTIPLE OBJECT OUTPUT

In order to properly assess the performance, we conducted various object tests. The results shown below show that yolo v3 is able to recognise every object that is there. When we examine the timings of algorithm for single and many items,

we find that multiple object detection takes longer and does excellent performance in detecting the far away objects as well.

Below is the performance of algorithm with multiple objects. In the Fig. 8, As a part of multiple detection, a snapshot of the image is taken with a person holding cell phone, the two objects person, cell phone are identified and voice message is prompted. In Fig. 9, An image is taken with two different objects bicycle, dog and the output generated is the audio as bicycle, dog. Similarly, in Fig. 10, feeding the input image with four multiple objects person, person, laptop and cell phone and the system gives the audio feedback of the objects identified.

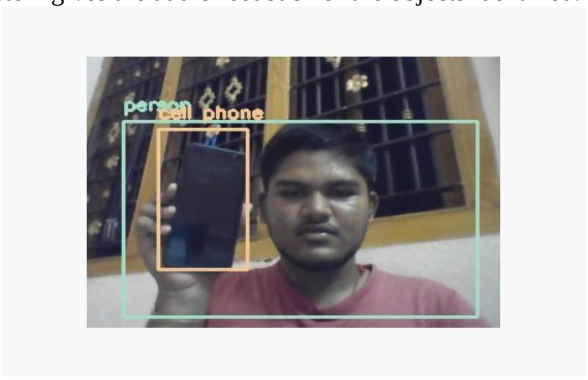


**Fig 8.An image is taken by the user holding a cell phone where the yolo_v3 detects the two objects and prompts a voice message as person, cell phone**


**Fig 9.An image is taken with two different objects as shown and the output generated is the audio as bicycle, dog**

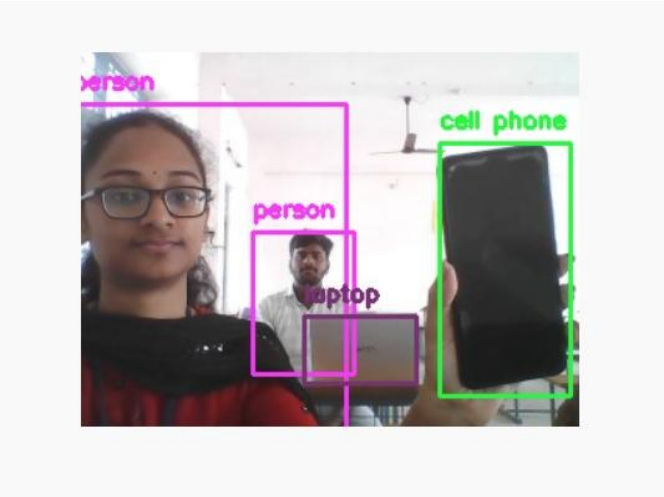


**Fig 10.In the multiple object detection, the image is given as input with four different objects and the audio feedback is person, person, laptop, cell phone**

Yolo_v3 is showing good accuracy in single object detection and multiple objects detection even if the objects are far and small in size. The key concern behind this contribution is to investigate the possibility of expanding the support given to the visually impaired people. It has been determined that YOLO may be used in conjunction with any IOT technology to advance a business.

## CONCLUSION AND FUTURE SCOPE

In order to recognise and categorise every item in front of a webcam with high accuracy and performance, YOLO_V3 can be appropriately used. We discover that YOLO_v3 is significantly more effective in recognising small items and distant objects to the test in a variety of scenarios. In this work, the input is tested for different features given by the user on the model. We are able to detect objects using yolo_v3 and use the web speech API to generate speech.YOLO_V3 is one of the fastest real-time object detection algorithm as compared to previous detection techniques. As soon as the snapshot of the image is taken within no time the output is generated. It is capable of processing 155 frames per second.

This algorithm can also be used in various ISM i.e., Industrial, Scientific and Medical applications such as face recognition like the system can be trained to detect the person by face and generate the output as the details of that person. Hence, the face recognition can be used in banking systems, automobile security, access control. The object detection technique can be applied in shopping malls, convenience stores, restaurants to detect people and store items. This helps in better understanding of shopper behaviour and improve operations. Object detection is a powerful tool for tracking and counting and tally how many people enter and exit a store.

One of the applications can also be the pedestrian detection where if there is any presence of object, it alerts the pedestrains with a voice prompt. Such systems are vital for detecting things like signs, other vehicles and lane markers. Precision and accuracy are parameters for manufactured products and goods. Object detection can identify parts and finished products that do not meet the quality standards. The technique can also be implemented in health risk assessment and drug discovery. In this study, we have done a small-scale evaluation of algorithm in terms of accuracy and preciseness. Future evaluations of more algorithms with more parameters and photos will be our goal. This time, we worked with a ready-made dataset, but moving forward, we'll aim to apply this technique to a more objective, customised dataset.

## ACKNOWLEDGMENT

We are grateful to our guide Professor Dr. P. Sudhakara Reddy for this continuous support and guidance. Through his guidance, we were able to successfully complete our project. Our sincere thanks go to our Head of the Department of Computer Science and Engineering at AITS Rajampet, for his support and time.

## REFERENCES

[1]. 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).
[2]. 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).
[3]. J. Redmon & A. Farhadi, “Yolov3: An incremental improvement,” ArXiv preprint arXiv: 1804.02767, (2018).
[4]. 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).
[5]. J. Redmon, & A. Farhadi, “YOLO9000: better, faster, stronger,” In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271) (2017).
[6]. 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).
[7]. 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).
[8]. 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).
[9]. S. Cherian, & C. Singh, “Real Time Implementation of Object Tracking Through webcam,” Internation Journal of Research in Engineering and Technology, 128-132, (2014)
[10]. 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).
[11]. 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).