International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 7 ǁ July 2022 ǁ PP. 508-514

"Real-time Object Detection using Deep Learning for helping People with Visual Impairments"

Ramyashree C A,
DepartmentofComputerScienceandEngineering,NitteMeenakshiInstituteoftechnology
DrSarojadeviH,
Professor and Head, Dept of CSE, NMIT, Bengaluru

Abstract --*For a livelihood, eyesight is among the most important senses. Millions and millions of people globally suffer with certain form of visual impairment. These people find it difficult to communicate and acquire information, which makes it challenging for them both to navigate safely and independently. The proposed work aims at transforming the visual world into an aural one by alerting blind individuals to the objects that are in their way. This will enable those with vision impairment to move autonomously without the need for outside aid by employing a real-time object detection system. This paper uses image processing and deep learning techniques to determine real-time objects through the webcam/laptop camera and inform blind people about the object through the audio output. The major goal of the proposed work is to have the visually impaired with good precision, the best performance outcomes, and a practical choice to improve their quality of life.*
Keywords: *Real time object detection, visually impaired, YOLO.*

Date of Submission: 05-07-2022 Date of acceptance: 19-07-2022

I. INTRODUCTION

Object detection is an issue in computer vision which is both technically hard and functionally useful. Detection is the task of identifying specific objects inside of the image. For a livelihood, eyesight is one of the most crucial senses. Millions of people around battle with the some means of visual impairment. These people find it hard to communicate and obtain information, which makes it difficult for them to function safely and independently. By alerting the blind to the items in their path, the proposed work aims to convert the visible world into an aural one. By utilizing the real-time object detection technology, this will aid those with vision impairment in freely navigating without the need for outside support. Though, in past few years, technology has made significant strides for people who are blind. Hands-free technology relies solely on human auditory input.

They benefit from not needing to interact visually or physically, which is advantageous for them. They can use screen readers to assist them in reading screens on gadgets. These object detection algorithms can be learned from start or they can be pre-trained. The majority of the time, we fine-tune pre-trained weights from pre-trained models in accordance with our needs and various use cases We can recognize and localize specific objects in an image or video using an object detection method that is based on vision. Object detection can be used to count the items in a scene, identify and locate them, and label them correctly using this method of localization and identification. By using object detection, we can simultaneously classify the different items we've discovered and find examples of those objects in the image. The computer vision techniques of image recognition and picture segmentation are closely related to that of object detection.

In this paper object detection is done for visually impaired persons we capture the video using webcam are using COCO dataset which contains ninety objects like person, traffic light, dog, cat, pen, car, bus etc. for pre training and here we use models YOLO V4 to object detection and Faster R-CNN algorithm to detect the object accurately and SSD-Mobilenet V2 is used for multiple object detection and the speech and the audio output is given to object detected.

II. MOTIVATION

The goal of this project is to provide people with vision impairments with the tools they need to freely navigate and locate objects in a new indoor environment and thoroughly learning a number of crucial skills, including pattern recognition and image processing.

III. OBJECTIVES

To recognize instances of a predefined set of object classes (e.g., people, cars, bikes, animals) and describe the locations of each detected object in the image using a bounding box. To develop an object detection and classification method. The process of detection and classification multiple objects using Realtime video

input feed.

To convert the recognized objects into speech which would help People with Visual Impairment.

IV. PROBLEMSTATEMENT

Input is the Real time video using web camera. Object detection, segmentation, localization and recognition using YOLO v4. To study and approach for object detection using Faster R-CNN to get accuracy. Detect multiple object classes from Realtime video input feed using SSD Mobile Net v2. Output is to detection accuracy is usually measured on a given test set where the expected outcome for a detection sample is compared to the actual outcome of the object detection system. The detection accuracy is the percentage of samples for which the expected outcome matches the actual outcome of the detection system along with speech output

V. PROPOSEDSYSTEM

In this paper object detection is done for visually impaired persons we capture the video using webcam Then the pre trained models are loadedand by using MS COCO dataset which contains ninety objects like person, traffic light, dog, cat, pen, car, bus etc. for pre training and here we use models YOLO V4 to object detection and Faster R-CNN algorithm to detect the object accurately and SSD-Mobile net V2 is used for multiple object detection ,then label the objects of MS COCO dataset and compare with objects that are captured real time objects and draw the bounding boxes around it and finally the object is detected. The detected objects are converted to speechas the audio output is given to object detected. The below figure 1 represents the proposed architecture of the system.

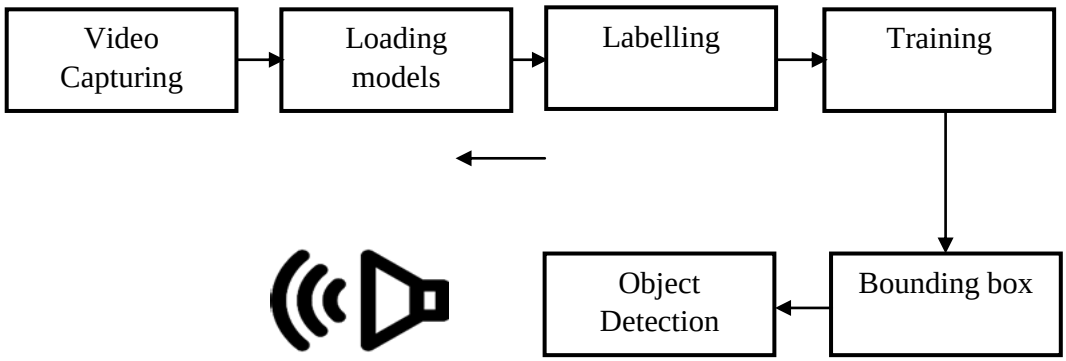

Fig 1: Proposed system architecture.

VI. DATASET DETAILS

In our work we have used MS COCO dataset, which MS COCO (Microsoft Common Objects in Context) is a dataset collection which has 90 or more image dataset persisting images of objects used on daily basis. The dataset contains annotations that can be used to train machine learning models to recognize, label, and describe objects. The objects are listed below in the table as shown in figure 2.

person	elephant	dog	skateboard	tv
bicycle	stop sign	horse	surfboard	toaster
car	bear	sheep	tennis racket	sink
motorcycle	parking meter	cow	bottle	refrigerator
airplane	zebra	tie	wine glass	book
bus	giraffe	skis	cup	clock
train	backpack.	snowboard	fork	vase
truck	umbrella	sports ball	knife	scissors
boat	handbag	kite	spoon	teddy bear
traffic light	tie	baseball bat	bowl	hair drier
fire	suitcase	baseball glove	banana	broccoli
couch	laptop	pizza	apple	carrot
potted plant	mouse	donut	sandwich	hot dog
bed	remote	cake	orange	cell phone
dining table	keyboard	chair	oven	microwave
toilet			toothbrush	

Fig 2:Table showing details of datasets

VII. IMPLEMENTATION

Firstly we set up the environment by install the following softwares like**:**

1.Python3.11
2. TensorFlow
3. Tensorboard
4.Mini Conda
5.YOLO V4
6.Mobilenet_SSD V2
7.Faster R-CNN

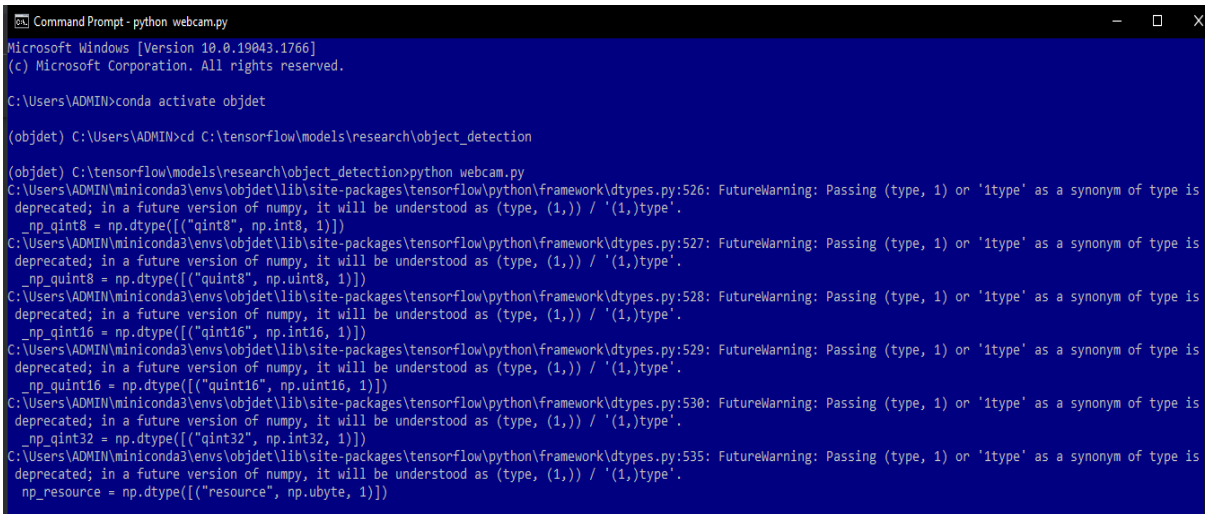

Fig 3: Screenshot of model being loaded loaded

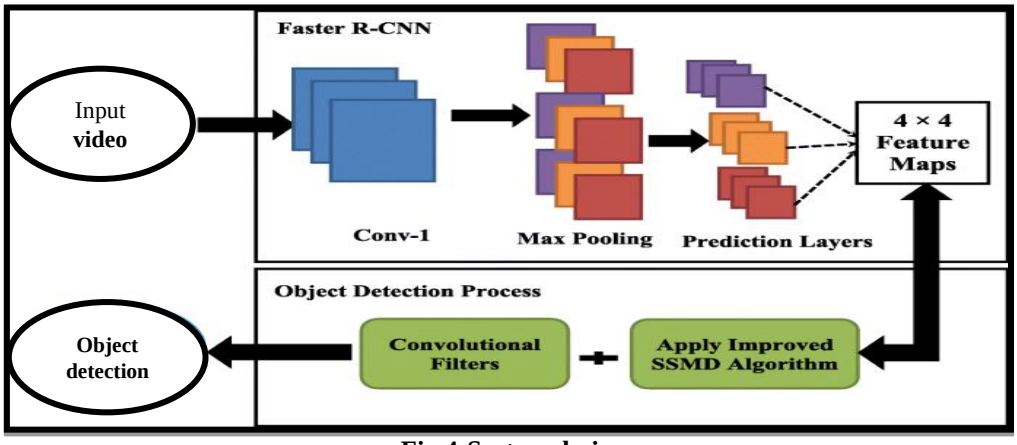

Fig 4:System design

The above figure represents the system design of the work.

1. YOLO V4

Theseobjectdetectionalgorithmscaneitherbelearnedfromscratchorwithpre-traineddata. In thebulkof use cases, we leverage pre-trainedweights from pre-trained modelsandthenadjustthemtoourspecificationsandvarioususecases.

You Only Look Once, or YOLO, is the abbreviation for a single shot recognition method. It is the perfect illustration of this method since it predicts classes and bounding boxes for the

entire image in a single run of the algorithm. The YOLO algorithm is essential for the reasons listed below.

- Speed: Because this algorithm can predict objects in real-time, it increases the speed of detection.
- High accuracy: The YOLO prediction method yields precise findings with few background mistakes.

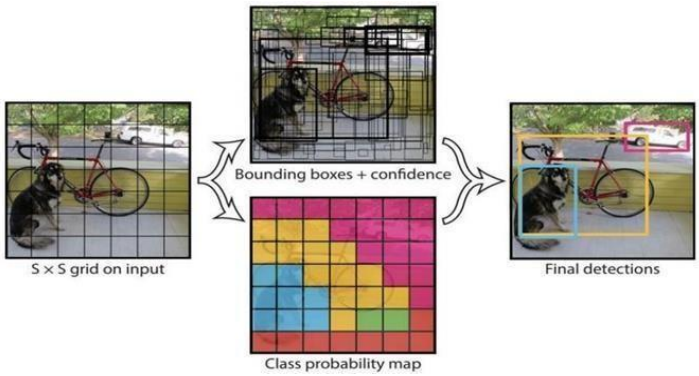

Fig5:RepresentationofYOLOAlgorithm.

Above figure 5 represents the Yolo algorithm. YOLO works by taking an image andsplitting it into an SxS grid, within each of the grid we take m bounding boxes. For each of thebounding boxes, the network gives an output a class probability and offset values for thebounding box. The bounding boxes have the class probability above which a threshold value isselectedandusedtolocatetheobjectwithintheimage.YOLOisordersofmagnitudefaster(45frames per second) than any other objectdetection algorithms.

The limitation of the YOLO algorithm is that it struggles with the small objects within the image, for example, it might have difficulties in identifying a flock of birds. This is due to the spatial constraints of the algorithm. The goal of object recognition is to identify every instance a recognized class of items in an image, including such people, cars, or faces. Even though there are often few instances of the object in the photograph, there are a vast array of locations and scales where they might appear that must be looked into in some way. Each image detection is supplied together with some kind of pose data.

This is about as simple to comprehend as the object's position, size, or extension as defined by its bounding box. Other times, the pose details are more detailed and also contain the parameters of the linear or non-linear transformation. As an illustration, face detection in a face detector may compute the locations of the eyes, nose, and mouth in addition to computing the bounding box of the face. The example of bicycle detection in a picture that pinpoints the locations of particular parts. The stance can also be expressed as a three-dimensional translation that shows the object's location in relative to the camera.
Convolutional slide window implementation Let's explore how to convert the network's fully connected layers into convolution layer before we explain how to create the sliding window utilizing convents.

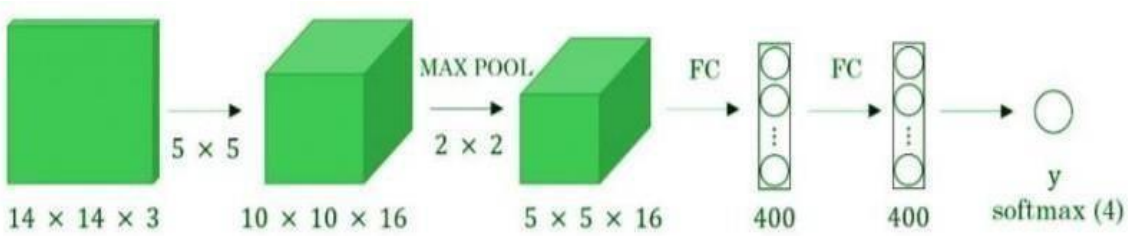

Fig 6: A straight forward convolutional network for each layer having two fully connected layers.

Above figure 6 represents a straightforward convolutional network for each layer having two fully connected layers. The extent of the object as specified by its bounding box is just as simple to comprehend as the object's position, scale, or other characteristics. Other systems allow for the conversion of a fully connected layer into a convolutional layer using a 1D convolutional layer. The dimensions of this layer match those of the fully linked layer in terms of height, width, and the number of filters.

Our project may use the idea of converting a fully connected layer into a convolution layer to enhance the model by substituting the fully connected layer with a 1-D convolution layers. The number of filters in the 1D convolution layers is identical to the geometry of the fully linked layer.

A convolution with the shape (1, 1, 4), which equals the maximum number of predicted classes, is also produced by the SoftMax layer. Let's continue to design a convolution iteration of the sliding window based on the aforementioned approach. Looking at the ConvNet that we trained to be in the following representation would help get us began (no fully connected layers). Assume the input image has had the dimension of 16 by 16 by 3 pixels size.

The sliding window method would have required us to submit the image to a aforementioned ConvNet four times, with each submission causing the sliding window to crop one portion of the input image matrix of size 14 14 3 and send it to the ConvNet. Instead, we instantly input the complete image (with the shape 16 x 16 x 3), which results in an output matrix with the shape 2 x 2 x 4 as a result. Each pixel in the output vector corresponds to a classification value for the cropped image and a probable crop result.

The outcome of the first sliding window, for instance, is represented by the left cell of the output matrix. The outcomes of the other cells in the matrix are the stride of the sliding window is decided by the number of filters used in the Max Pool layer. In the example above, the Max Pool layer has two filters, and for the result, the sliding window moves with a stride of two resulting in four possible outputs to the given input. The main advantage of using this technique is that the sliding window runs and computes all values simultaneously.

2.Faster R-CNN

The main idea is to use shared convolutional layers for region proposal generation and for detection. The authors discovered that feature maps generated by object detection networks can also be used to generate the region proposals. The fully convolutional part of the Faster R-CNN network that generates the feature proposals is called a region proposal network (RPN). The authors used Fast R-CNN architecture for the detection network.

By switching between training for RoI generation and detection, a Faster R-CNN network is created. Two distinct networks are first trained. These networks are then integrated and adjusted. While fine-tuning, some layers are maintained fixed while others are trained one at a time. One image is fed into the trained network as input. Using the image, the shared fully convolutional layers produce feature maps. The RPN receives these feature maps. The RPN generates region suggestions, which are fed into the final detection layers along with the aforementioned feature maps. These layers produce the final classifications and contain a RoI pooling layer.

Region proposals are computationally almost free when using shared convolutional layers. The fact that a GPU may be used to compute the region proposals on a CNN is an additional advantage. Utilizing a CPU, traditional RoI generating techniques like Selective Search are applied. Instead of employing a pyramid of scaled images or a pyramid of varied filter sizes to deal with the detection window's various shapes and sizes, the method uses customised anchor boxes. As reference points for several region suggestions centred on the same pixel, the anchor boxes serve this purpose.

3.Mobile SSD

The Single Shot MultiBox Detector (SSD) takes integrated detection even further. The method does not generate proposals at all, nor does it involve any resampling of image segments. It generates object detections using a single pass of a convolutional network. Somewhat resembling a sliding window method, the algorithm begins with a default set of bounding boxes. These include different aspect ratios and scales. The object predictions calculated for these boxes include o set parameters, which predict how much the correct bounding box surrounding the object identifiers from the default box.

The algorithm deals with different scales by using feature maps from many different convolutional layers (i.e. larger and smaller feature maps) as input to the classier. Since the method generates a dense set of bounding boxes, the classier is followed by a non- maximum suppression stage that eliminates most boxes below a certain confidence threshold.

VIII. Results

Firstly, the transfer learning method was used. This method takes the weights andknowledge of an already existing model and applies it to a different one which should deal withsimilar problems. Images and bounding box data was downloaded through the API of MSCOCOwhich is a dataset providing more than 2.5 million label instances in 328 thousand images. [Thereare other datasets such as Pascal VOC and ImageNet but none of them provided an easy way todownloaddatasimi- lartoMSCOCO.For the followingmodels 9classes is downloadedwithone thousand images and labels each to test, vary and experimentwith the amount of datapro-videdtothemodel.

Secondly,in this work we havedownloadedthemodelsfromTensorflow,onewasthefaster-rcnn-inception-v2-coco

and the ssd-mobilnet-v2-coco trained with the dataset from MSCOCO. These are 2 modelswherethefirstoneisimplementedwith the Faster-RCNN and the later with SSD algorithm.
After running the program, a new window will open, which can be used to detect objects in real time. We trained using a library of visuals generated by Coco for more than 80 different subjects, and we got overall success rates (right classifications on choice) of 99.6%, 98.6%, 97.4%, and 97%, respectively.

Fig 7: Snapshot of result 1

The number of key features provided to the algorithm and the size of the database affect how quickly an object can be identified in a more-or-less linear fashion. Currently, the total recognition times for the 6-object database and the 8-object information on a single CPU were about 20 and 2 seconds, respectively. By advancing the simulated and measured, this could also be significantly improved. An average of 2–4 cycles per second (FPS) are refreshed by the application into the video window each 0.25–0.5 seconds. With the use of a camera, we detect live items in this project. It describes me as having 95 percent confidence, and it also characterizes the water bottle as having 95 percent confidence. It illustrates how completely the object was detected.
The object identification using YOLOis shownbelowusing pre-trainedYOLOweights, whichare available on the officialDarknet repository of YOLO,with examples andcomprehensivedocumentation. At the time of writing, there are four primary varieties of the strategy: YOLOv1,YOLOv2,YOLOv3,andYOLOv4.
The first version offered a generic architecture; the second iterated the design and improved thebounding box proposal by using specified anchor boxes; the third iterated the model architectureandtrainingprocedure.TheYOLOV4modelwasutilizedinthisproject.

Fig 8: Snapshot of result 2

Finally, when an object is detected, its name is displayed including the speech that identifies it. The object that is being detected and its confidence level will be displayed.

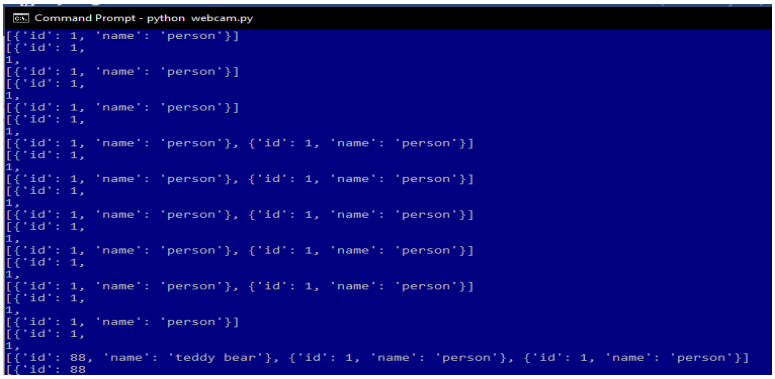

Fig 9: Sample output.

IX. CONCLUSION AND FUTURE ENHANCEMENT.

The primary issue with the earlier approaches was how to restore the precision loss, which SSD addressed with various enhancements including multi-scale feature maps and default boxes.

Feature maps are utilised for small item detection at greater resolutions. The three primary components of the training set for the improved SSD method are choosing the box size, box matching, and loss function. This algorithm converts detected items into speech to aid people with visual impairments.

The Object Detection system in Images is web-based application which mainly aims to detect the multiple objects from various types of images. To achieve this goal shape and edge feature from image is extracted. It uses large image database for correct object detection and recognition. This system will provide easy user interface to retrieve the desired images. The system has additional feature such as Sketch based detection. In Sketch detection user can draw the sketch by hand as an input. Finally, the system results output images by searching those images that user want.

REFERENCES

[1] "An improved method for visual surveillance using background subtraction technique". 2nd International Conference on Signal Processing and Integrated Networks", Sharma, L. S. and Yadav, D. K.,.2018 IEEE, February, pp. 421-426. "Detection of Moving Objects Using Fuzzy Color Difference Histogram Based Background Subtraction. Signal Processing Letters". Panda, D. K. and Meher, S., 2018., IEEE, January, Vol. 23, Issue 1, pp. 45-4.

[2] "Detection of Dynamic Background Due to Swaying Movements from Motion Features". Pham, D. S., Arandjelovic, O., and Venkatesh, S. 2019 IEEE Transaction on Image Processing, January, Vol. 24, Issue 1, pp. 332–344.

[3] "A Robust Adaptive Algorithm of Moving Object Detection for Video Surveillance". Kermani, E. and Asemani, D., 2019.IEEE, May.

[4] "Background Subtraction using Illumination Invariant Structural Complexity". Kim, W. and Kim, Y., 2020. IEEE Signal Processing Letters, March, Vol. 23, Issue 5, pp 634-638.

[5] "Detection of Moving Objects Using Foreground Detector and Improved Morphological Filter". Olugboja, A. and Wang, Z., 2020. November, pp. 329 – 333.

[6] "An Optimised Background Modelling for Efficient Foreground Extraction". Sivagami, M., Revathi, T., and Jeganathan, L. 2020. IEEE March, Vol. 10, Issue 1/2, pp.44-53.

[7] "Real-time Object Detection with Deep Learning for Robot Vision on Mixed Reality Device". Jiazhen Guo, Peng Chen, Yinlai Jiang. 2021. IEEE May, Vol. 11, pp.44-53.

[8] "Real-time Object Detection for helping People with Visual Impairments.MatteoTerreran , Andrea G. Tramontano, Jacobus C. Lock.2021, IEEE Transaction on Image Processing, January, Vol. 24, Issue 1, pp. 332–344.

[9] "Adaptive pedestrian Detection in Infrared Images using Fuzzy Enhancement and Top-Hat Transform". Soundrapandiyan, R. and Mouli, P. V. S. S. R. C. 2021 International Journal of Computational Vision and Robotics, Inderscience, January, Vol. 7, Issue 1/2, pp. 49-6.