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 Voice Feedback and Distance Estimation

Swarang Dani, Rutuja Gadhave, Abhiraj Bandal, Nikita Kamble
Department of Computer Engineering, Zeal College of Engineering and Research, Pune, Maharashtra, India

Article Info
Volume 8, Issue 3
Page Number : 503-510

Publication Issue :
May-June-2022

Article History
Accepted: 10 May 2022
Published: 30 May 2022

ABSTRACT

The rise of visual impairment is increasing worldwide. There are many people who have problems of visual impairment and blindness. A person with little or no vision usually relies on walking sticks, smart glasses, braille systems and other optical tools. However, to assist those who have near-sightedness or no vision at all, the visual world must be converted into an audio world, which can help them understand surrounding objects and their spatial locations. As a result, we propose to assist the visually impaired by providing a system that is the most practicable, simplistic and cost-effective. Object detection is an important technique based on computer vision that identifies objects in images and videos. In this project, we are detecting objects in real time, using You Only Look Once v3 (YOLO v3) which is a real time object detection algorithm that runs on a variation of an extremely complex Convolutional Neural Network (CNN) architecture called the Darknet, after object is detected, we estimate the distance from camera to object and using a text-to-speech conversion tool like gTTS, provide voice feedback to the user.

Keywords— Object Detection, YOLO, Darknet, gTTS, Deep neural network, TensorFlow, OpenCV

I. INTRODUCTION

Vision Impairment or blindness is a problem faced by many people of different ages, it is a condition where a person has a decreased ability to see, and it can be caused due to cataracts, trachoma, glaucoma, corneal opacity or, age-related macular degeneration. People who are blind or visually impaired find it difficult to navigate around and there are many difficulties faced by them in their day-to-day living. Finding information about objects and having difficulty in navigating around are two major challenges that a blind person must face. Simple items are difficult for them to distinguish, and objects with similar forms are even more challenging. Blind people usually depend on white canes, sticks or people who help them in walking. Since they do not have a sense of sight, it becomes difficult for them to detect sudden changes in the environment. With the recent advancements in Computer Science Technology (CS), much research has been conducted to solve inconveniences in daily life, and as a result, various solutions for people have been provided such as eye-ring project, smart glasses, recognition systems, electronic mobility aids and so on. But the drawbacks of these solutions are they are very expensive, heavy

and less robust. In order to overcome these drawbacks, we have proposed a system that is built using advanced state-of-the-art technologies such as image-processing and deep learning.

The proposed solution makes use of a web camera to capture real time images, which are then passed as input to the system. In order to improve the quality and remove undesired distortions, image pre-processing is done, the background and foreground are then separated and a deep neural network module with a pre-trained YOLO V3 model is applied resulting in feature extraction. To recognize the object in the image we match known object features with the extracted features. After the object is successfully detected, the name of the object along with its distance is conveyed to the person through voice output.

This system is aimed at developing an affordable, robust, light-weight and easy to use application, that will help the visually impaired person perform his daily tasks with ease.

II. LITERATURE SURVEY

Unique Smart Eye Glass for Visually Impaired

This paper proposes a system that helps the blind to detect obstacles. This system makes use of ultrasonic sensor for distance measurement and a microcontroller. A GSM SIM900A module is used for notifying the guardian of blind person about time, temperature and location through SMS. This system fails to detect ground level objects and becomes heavy due to extra hardware components [1]

Reader and Object Detector for blind

This paper proposes a system that helps visually impaired people by detecting objects nearby and assisting them in reading text material. The Implementation is done by making use of microcontroller such as Raspberry Pi and text reading is supported by a software named OCR. TTS Synthesis is used to convert text format into audio format, Object Detection is performed using TensorFlow API. This system can only identify words of English Language and it can read words that have font size of 14. The extra hardware makes this system expensive and heavy [2]

An Embedded Real-Time Object Detection and Measurement of its Size

In this paper objects and their sizes are detected in real-time video streams using OpenCV as a software library, Raspberry Camera and Raspberry Pi 3 model. This system thus recognizes and measures the size but this paper doesn't provide the location of the object [3]

Object Recognition App for Visually Impaired

This paper proposes an android application to help blind people see through handheld device like mobile phone. A Single Shot Detector (SSD) algorithm is used for object detection and voice output is provided through a text-to-speech API. COCO Dataset is used which has 80 different object classes with 83k Training Images and 41k Testing Images [4]

III. PROPOSED WORK

The proposed system consists of the following three parts, they are object detection, text-to-speech conversion and estimation of distance from camera to object.

A. Object Detection:

Object Detection is a technique based on computer vision that aims at locating and classifying objects in images or videos. There exist several deep learning-based methods for object detection such as R-CNN [5], Fast R-CNN [6], Faster R-CNN [7], SSD [8] and YOLO [9]. However, we need to choose an algorithm that works efficiently and gives better accuracy and speed as compared to other algorithms.

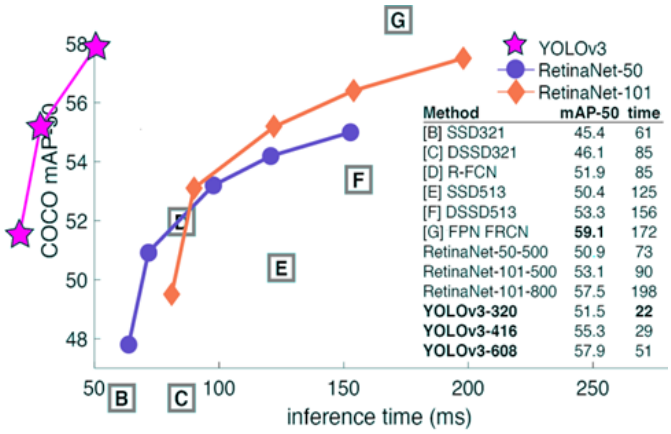


Figure 1: Performance of Object Detection Algorithms

The above figure shows a comparison of different object detection algorithms. All these algorithms have been trained on a COCO Dataset which is a large-scale object detection dataset containing a total of 80 object categories. YOLO v3 is extremely fast and accurate with a mAP of 51.5% and an inference time of 22 milliseconds.

For this system, YOU ONLY LOOK ONCE (YOLO) algorithm is used. YOLO has several advantages over other algorithms. It requires a single run through a network to detect objects when contrasted with other algorithms like R-CNN which require thousands for a single picture. This makes it incredibly quick, in excess of multiple times quicker than R-CNN.

YOLO

YOLO abbreviated as "You Only Look Once" is a 1-stage detector algorithm that makes use of a convolutional neural network to detect objects in real time. It was first proposed by Joseph Redmon, Santosh Divvala, Ross Girshick and Ali Farhadi and is a popular algorithm because of its high speed and accuracy.

In YOLO, the input image is first divided into various grids of size S x S. Each grid cell predicts B Bounding boxes and their confidence scores. Confidence score indicates whether an object is present in the box and how accurate the box is. Confidence score is defined as:

$$C = Pr\ (Object) * IOU$$

where IOU stands for intersection over union and has a value of either 1 or 0. A value of 0 indicates there is no overlap between the boxes and value of 1 indicates they are completely overlapping

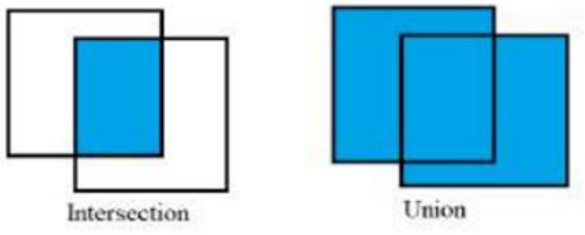


Figure 2: Illustration depicting intersection and union

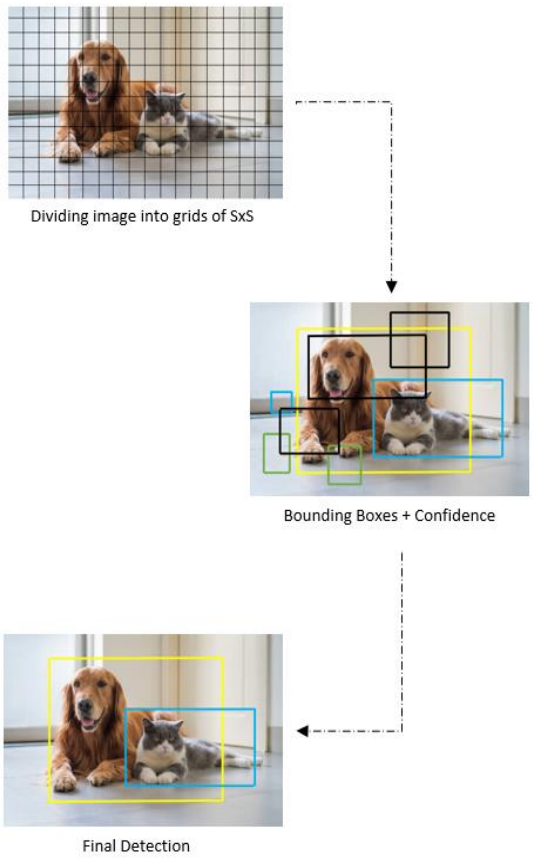
Once the bounding boxes are made, each bounding box will contain certain grid cells that predict C conditional class probabilities. We multiply these class probabilities with individual box confidence predictions to get class-specific confidence scores for each box.

$$Pr\ (Class_i \mid Object) * Pr\ (Object) * IOU$$
$$= Pr\ (Class_i) * IOU$$

YOLO uses the following equation to calculate loss –

$$\begin{aligned} Loss = & \\ & \lambda_{coord} \sum_{i=0}^{s^2} \sum_{j=0}^{A} \mathbb{1}_{ij}^{obj} [(b_{x_i} - b_{\hat{x}_i})^2 + (b_{y_i} - b_{\hat{y}_i})^2] \\ & + \lambda_{coord} \sum_{i=0}^{s^2} \sum_{j=0}^{A} \mathbb{1}_{ij}^{obj} [(\sqrt{b_{w_i}} - \sqrt{b_{\hat{w}_i}})^2 + (\sqrt{b_{h_i}} - \sqrt{b_{\hat{h}_i}})^2] \\ & + \sum_{i=0}^{s^2} \sum_{j=0}^{A} \mathbb{1}_{ij}^{obj} (C_i - \hat{C}_i)^2 \\ & + \lambda_{noobj} \sum_{i=0}^{s^2} \sum_{j=0}^{A} \mathbb{1}_{ij}^{noobj} (C_i - \hat{C}_i)^2 \\ & + \sum_{i=0}^{s^2} \mathbb{1}_{i}^{obj} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2. \end{aligned}$$

The λ_{coord} increases the weight for the loss in the boundary box co-ordinates and λ_{noobj} weights down the loss when detecting background. Here C refers to the confidence score, and $p_i(c)$ denotes the conditional class probability for class c in cell i. The performance of a model depends on the loss, higher the loss, lesser the performance and vice versa



The above image shows the three techniques that are applied to produce final detection results. The accuracy of predictions made by models in object detection are calculated through the average precision equation shown below.

$$\text{avg}\,Precision = \sum_{k=1}^{n} P(k)\,\Delta r(k)$$

where P(k) refers to the precision at threshold k and Δr(k) refers to change in recall.

B. Text-to-Speech Conversion:

Text-to-speech also known as speech synthesis is a technology in which normal language text is taken as input and converted into speech as output. It involves many disciplines such as acoustics, digital signal processing and statistics. A TTS conversion tool can be used to assist the blind, for example the contents on a screen can be automatically read aloud to the user. A blind person has no vision, but they can learn to interpret visual input by making use of their ears, it gives them some idea of what objects are present in their surroundings.

For this system, Google text-to-speech (gTTS) tool is used, it is a free and easy to use API and supports several languages including English, Hindi, French, German, etc. gTTS has loads of features such as customizable speech-specific sentence tokenizer that allows to read unlimited lengths of text, and customizable text pre-processors that provide corrections in pronunciations.

C. Distance Estimation

Distance estimation is important as it allows us to find the future location of an object based on the current distance. There exist many methods of finding distance from camera to object such as laser technology, stereo vision, ultrasonic sensors and so on, but all these methods have certain drawbacks, they are expensive, heavy and can get affected from external environment. We don't always need to make a system heavy by adding unnecessary hardware modules. Since we already have an integrated camera for our object detection, we can make use of it to estimate the distance.

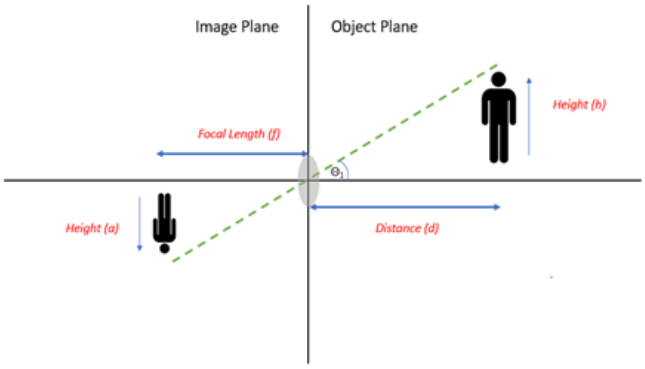


Figure 3: Principle of Similar Triangles

In order to determine the distance from camera, we find out the actual distance(d) and actual height(h) or width(w) of the object. Since it is not possible to capture entire object from toe to head in the camera, for this reason we measure width of the object rather than the height.

After getting the actual measurements we capture the reference image and find out the apparent width(a) of the object that is detected in camera Through this we can get the focal length(f) of our camera. Using principle of similar triangles figure 3, we can obtain the formula as follows:

$$\frac{d}{f} = \frac{h}{a}$$

$$f = \frac{d \times a}{h} \quad \text{pixels}$$

Since focal length and the actual object width are now constants, if we move camera both closer and farther away from the object, we can determine the distance of the object.

$$d = \frac{f \times h}{a}$$

where,

f = Focal length

d = Distance from object to camera lens

a = Detected object width

h = Actual object width

IV. IMPLEMENTATION

A. Description:

The proposed system uses Python as a programming Language. It is simple and easy to use and is most preferred language because of its clear syntax and readability. It has a wide variety of machine learning frameworks and libraries and is mostly used for producing deep learning algorithms.

OpenCV is another great library that carries out computer vision tasks and is used for image processing. This library has more than 2000 algorithms that includes machine learning, computer vision and image processing algorithms.

NumPy also called Numerical Python is a library in python for data manipulation and scientific computing. It is used to perform wide variety of mathematical operations on multi-dimensional arrays. TensorFlow is an amazing open-source library used for deep learning applications. It supports both CPU and GPU computing, and is much faster than other deep learning libraries like Keras and PyTorch.

The system uses COCO dataset which is a popular and widely used dataset for machine learning and deep learning related work. It contains 80 object categories and over 300K images out of which 200k images are labelled.

B. System Architecture:

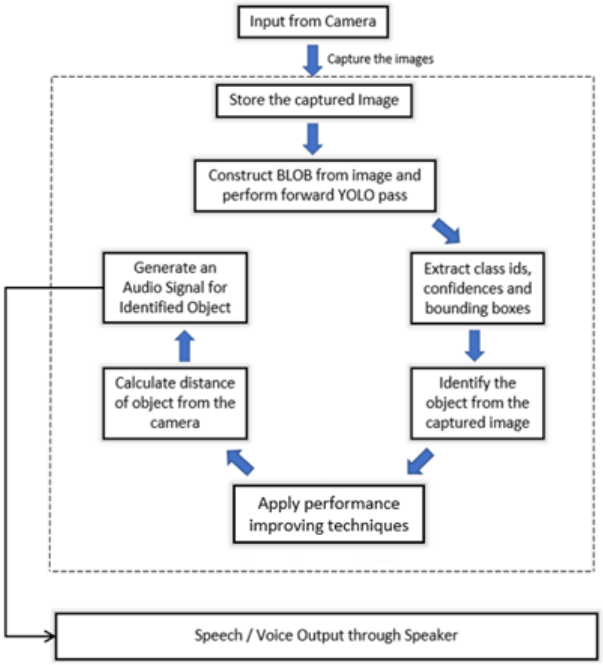


Figure 4: System Architecture

The above figure depicts the architecture of our system. A web camera starts capturing real time images, these images are stored and then passed as input to a computer-based system. To improve the quality and remove distortions from the image we perform Image pre-processing. Image pre-processing is a method that converts an image into digital form and performs operations on it, in order to get an enhanced image or to extract useful information from

it. Image pre-processing tasks normally involve Mean Subtraction and Scaling. Mean subtraction helps in levelling uneven sections of an image or detecting changes between two images, whereas Scaling refers to resizing the image. OpenCV provides a function called blob From Image that helps in performing tasks of image pre-processing. We now make use of our model for detecting objects in the image. The model divides every input image into grids of size SxS and each grid predicts B bounding boxes, along with the probability of the class which the object belongs to.

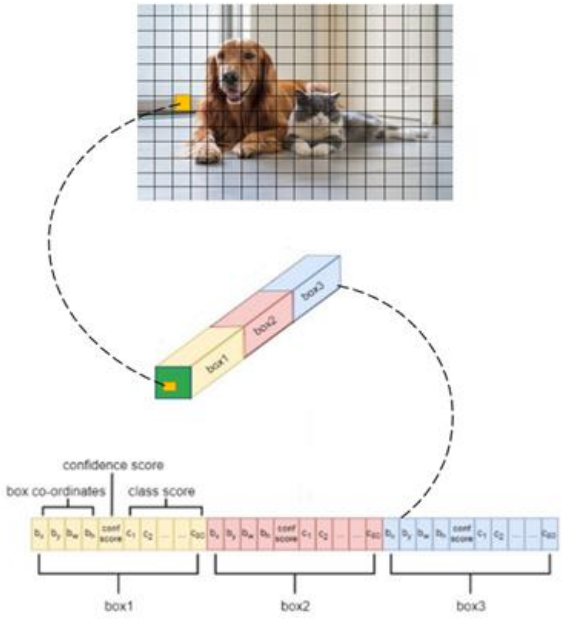


Figure 5: Prediction of an object

Each bounding box has 5 attributes they are center co-ordinates (b_x , b_y) , height (b_h) , width (b_w) and confidence score (c). Since there can be multiple bounding boxes, we perform non-max suppression and select the bounding box that has the highest confidence score. The same process is applied for remaining boxes until there are no more boxes left. Finally, we are able to detect object in the frame.

Once the object is detected we calculate the distance of the object from the camera. The detected objects and their absolute location are then notified to the blind person through a voice output.

V. RESULTS AND DISCUSSION

The experimental results prove that the proposed system can detect objects and notify the blind about nearby objects along with its distance. The system has a frame rate of 8 -10 FPS without GPU and takes 2000 ms of average computational time for detection. More than 80% of the objects are detected and recognized accurately. YOLO v3 has good accuracy and precision with a mAP of 51.5% and inference time of 22 milliseconds.

The model works very well, in order for the objects to be properly detected in the frame, they should not be too close to the camera frame and should be at a distance more than the focal length of the lens.

Figure 6: Detection of Monitor, Keyboard and Mouse

Figure 7: Detection of remote

VI. CONCLUSION AND FUTURE WORK

The proposed system identifies objects in real time and provides feedback to the user about nearby objects along with its distance. The system is able to detect objects with an average accuracy of 85%. This system assists the blind in his day-to-day activities and helps them to overcome the threats that they may come across in their daily life.

The distance of only few objects such as keyboard, mouse, bottle, cup, cellphone, etc can be estimated. It is not possible to calculate distance of large objects like train, airplane, boat, truck, etc, however this can be made possible by using advanced distance measurement tecnhniques like stereo-vision in the future stages. The accuracy of detection in darkness must be improved. In the COCO dataset, there are currently 80 object categories consisting of person, cellphone, keyboard, mouse, toothbrush, cat, dog, motorbike and many more, the system can only recognize the objects present in the COCO dataset, we are working on to include more objects in the dataset in future. Some additional features such as finding exact location of the user and notifying the user in which direction they should move can also be considered in the future stages.

VII. REFERENCES

[1]. M. R. Miah and M. S. Hussain, "A Unique Smart Eye Glass for Visually Impaired People," 2018 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE), 2018.

[2]. M. Murali, S. Sharma and N. Nagansure, "Reader and Object Detector for Blind," 2020 International Conference on Communication and Signal Processing (ICCSP), 2020.

[3]. N. A. Othman, M. U. Salur, M. Karakose and I. Aydin, "An Embedded Real-Time Object Detection and Measurement of its Size," 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), 2018.

[4]. S. A. Jakhete, P. Bagmar, A. Dorle, A. Rajurkar and P. Pimplikar, "Object Recognition App for Visually Impaired," 2019 IEEE Pune Section International Conference (PuneCon), 2019.

[5]. R. Girshick, J. Donahue, T. Darrell and J. Malik, "Region-Based Convolutional Networks for Accurate Object Detection and Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016.

[6]. R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), 2015.

[7]. S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[8]. Liu, W. et al. (2016). SSD: Single Shot Multibox Detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016.

[9]. Aparna Mote,Prajakta Markad , Satyam Takawale , Nikhil Lomate , "Emergency Care APP – Analysis and Review", International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) Volume 6, Issue 1, June 2021

[10]. J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[11]. A. Karthik, V. K. Raja and S. Prabakaran, "Voice Assistance for Visually Impaired People," 2018 International Conference on Communication, Computing and Internet of Things (IC3IoT), 2018.

[12]. Rahul and B. B. Nair, "Camera-Based Object Detection, Identification and Distance Estimation," 2018 2nd International Conference

on Micro-Electronics and Telecommunication Engineering (ICMETE), 2018.

[13]. N. Zhang and J. Fan, "A lightweight object detection algorithm based on YOLOv3 for vehicle and pedestrian detection," 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), 2021.