Object Detection System with Voice Output using Python

Rajeshwar Kumar Dewangan[1], Dr. Siddharth Chaubey[2]

[1]Research Scholar, [2]Assistant Professor
Department of Computer Science & Engineering,
Shrishankaracharya Group of Institute, Bhilai, C.G., India

***Abstract*: As object recognition technology has developed recently, various technologies have been applied to autonomous vehicles, robots, and industrial facilities. However, the benefits of these technologies are not reaching the visually impaired, who need it the most. In this paper, we proposed an object detection system for the blind using deep learning technologies. We use voice recognition technology in order to know what objects a blind person wants, and then to find the objects via object recognition. Furthermore, a voice guidance technique is used to inform sight impaired persons as to the location of objects. The object recognition deep learning model utilizes the Single Shot Multibox Detector (SSD) neural network architecture, and voice recognition is designed through speech-to-text (STT) technology. In addition, a voice announcement is synthesized using text-to-speech (TTS) to make it easier for the blind to get information about objects. The system is built using python OpenCV tool. As a result, we implement an efficient object-detection system that helps the blind find objects in a specific space without help from others, and the system is analyzed through experiments to verify performance.**

***Keywords*: Object detection, Voice output, Python OpenCV, CNN**

I. 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, there are still many inconveniences for the visually impaired. The greatest inconveniences 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. Previous studies included object analysis using ultrasonic sensors. However, with these methods, it is difficult to know exactly where an object is located, especially in the presence of obstacles. In this paper, we analyze accurate object information and obtain a location using a deep learning object recognition technique. In addition, voice recognition and voice guidance technologies are synthesized so the visually impaired can know the location of the objects they want to find by speaking to the system. Object recognition algorithms are designed based on theSingle Shot MultiBox Detector (SSD) structure, an object recognition deep learning model, to detect objects using a camera. In addition, voice recognition technology designed to use speech-to-text (STT) technology converts a user's vocal commands into text, from which only specific words are extracted and retrieved by the system. In the voice guidance technology, the technique of synthesizing the position of the article so it can be output, and synthesizing the name of the article, is done by using text-to-speech (TTS). In this paper, we propose an efficient object detection system to help find objects in a certain space without help from others, with special consideration for the blind.

1.1 Object Detection:

Object recognition has developed rapidly, starting with the deep learning–based convolutional neural network (CNN) technique [5] that drew attention at the ImageNet 2012 competition. The CNN, however, was accurate with object classification, but it was difficult to determine where inside the image the object was located. Subsequently, the model for solving this problem was the region-based consolidated neural network (R-CNN), which uses a linear regression method. However, due to the slow speed of the R-CNN, Fast R-CNN was developed. It utilizes a deep learning technique to not only classify the object but also to find the area the object is located in. Nonetheless, there was a limit in that the above model's object recognition processing speed was insufficient for real-time object recognition. Since then, You Only Look Once (YOLO), which comprises all the processes of object recognition as a deep learning network, has emerged, and technologies with fast detection speeds, such as Single Shot MultiBox Detector (SSD), have been developed. YOLO estimates the type and location of objects using regression inference on the problem of area selection and classification. On the other hand, SSD does not create candidate areas separately, but recognizes objects using a feature map of various sizes. Since it does not generate candidate areas, it is faster to train than the Faster R-CNN and is more accurate than YOLO because it uses different sizes of feature map.

1.2 Machine Learning: Machine learning is an application of artificial intelligence (AI) which provides a system which has the ability to automatically learn itself and improve from its experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

1.3 Image Processing: Image processing is a method to perform some operations on an image, in order to get an enhanced image or to extract some useful information from it. It is a type of signal processing in which input is an image and output may be image or characteristics/features associated with that image. Image processing basically includes the following three steps:

- Importing the image to the system.

- Analyzing and manipulating the image.
- Output in which result can be altered image or report that is based on image analysis.

The second section of the paper discloses about the previous studies similar to this project. Proposed system is explained in section third. Fourth section comprises of results and analysis. At the end, the conclusion of the paper is mentioned.

II. Proposed System

The system is implemented using OpenCV python which detects various objects in real-time along with real-time text reader.

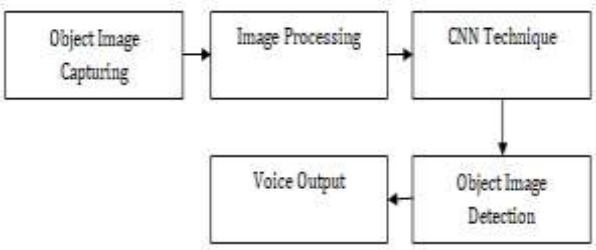

Figure 2.1: Flowchart

OpenCV library is used for image processing since it provides support to real time applications. Python programming language is used for build the machine learning model. TensorFlow library is used for writing machine learning application process. It provides high performance numerical computation. It has a flexible architecture which makes easy deployment of computation across a variety of platforms possible.

Object Information Extraction: A TensorFlow object detection application programming interface (API) based on an SSD deep learning structure was used to detect the location of objects. Because 90 objects had already been learned in this API, weights generated in the model were utilized in our model. In this study, the source code of the detection model that explores objects and draws boundaries was analyzed and modified because the location of objects must be identified in real time. The coordinates of the object's camera were extracted from the source code.Object detection is a computer vision technique that allows us to identify and locate objects in an image or video. With this kind of identification and localization, object detection can be used to count objects in a scene and determine and track their precise locations, all while accurately labeling them.

Convolution Neural Network: In neural networks, Convolutional neural network (ConvNets or CNNs) is one of the main categories to do images recognition, images classifications. Objects detections, recognition faces etc., are some of the areas where CNNs are widely used.CNN image classifications takes an input image, process it and classify it under certain categories (Eg., Dog, Cat, Tiger, Lion). Computers sees an input image as array of pixels and it depends on the image resolution.Technically, deep learning CNN models to train and test, each input image will pass it through a series of convolution layers with filters (Kernals), Pooling, fully connected layers (FC) and apply Softmax function to classify an object with probabilistic values between 0 and 1. Convolution is the first layer to extract features from an input image. Convolution preserves the relationship between pixels by learning image features using small squares of input data. It is a mathematical operation that takes two inputs such as image matrix and a filter or kernel.

Object detection techniques:

Single shot multibox detector: The single shot multibox detectoris one of the best detectors in terms of speed and accuracy comprising two main steps, feature map extraction and convolutional filter applications, to detect objects. The SSD architecture builds on the VGG-16 network, and this choice was made based on the strong performance in high-quality image classification tasks and the popularity of the network in problems where transfer learning is involved.

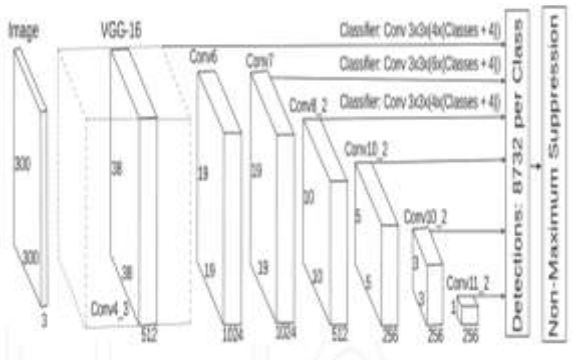

Figure 2.2: The SSD network has several feature layers to the end of the base network, which predicts the offsets to default boxes of different scales, aspect ratios, and their associated confidences.

Instead of the original VGG fully connected layers, a set of auxiliary convolutional layers change the model, thus enabling to extract features at multiple scales and progressively decrease the size of the input to each subsequent layer. The bounding box generation considers the application of matching precomputed, fixed-size bounding boxes called priors with the original distribution of ground truth boxes. These priors are selected to keep the intersection over union (IoU) ratio equal to or greater than 0:5.

$$L(x,c,l,g) = 1/N (L_{conf}(x,c) + \alpha L_{loc}(x,l,g))\text{………}(1)$$

where N is the number of matched default boxes and L_{conf} and L_{loc} are the confidence and location loss, respectively.

You only look once: You only look once is a state-of-the-art object detection algorithm which targets real-time applications, and unlike some of the competitors, it is not a traditional classifier purposed as an object detector. YOLO works by dividing the input image into a grid of S S cells, where each of these cells is responsible for five bounding boxes predictions that describe the rectangle around the object. It also outputs a confidence score, which is a measure of the certainty that an object was enclosed. Therefore the score does not have any relation with the kind of object present in the box, only with the box's shape. For each predicted bounding box, a class it's also predicted working just like a regular classifier giving resulting in a probability distribution over all the possible classes. The confidence score for the bounding box and the class prediction combines into one final score that specifies the probability for each box includes a specific type of object. Given these design choices, most of the boxes will have low confidence scores, so only the boxes whose final score is beyond a threshold are kept. Equation (2) states the loss function minimized by the training step in the YOLO algorithm.

$$\lambda_{coord} \sum\sum 1_{ij}^{obj} [(x_i-x'_i)^2 + (y_i - y'_i)^2] + \lambda_{coord} \sum\sum 1_{ij}^{obj} [(w_i^{1/2}-w'_i{}^{1/2})^2 + (h_i^{1/2} - h_i^{1/2})^2] + \lambda_{coord} \sum\sum 1_{ij}^{obj} [(c_i-c'_i)^2]\text{…………………………………....}(2)$$

where 1_{ij}^{obj} indicates if an object appears in cell i and 1_{ij}^{obj}denotes the j th bounding box predictor in cell i responsible for that prediction; x, y, w, h, and C denote the coordinates that represent the center of the box relative to the bounds of the grid cell. The width and height predictions are relative to the whole image. Finally, C denotes the confidence prediction, that is, the IoU between the predicted box and any ground truth box.

Faster region convolutional neural network: The faster region convolutional neural network is another state-of-the-art CNN-based deep learning object detection approach. In this architecture, the network takes the provided input image into a convolutional network which provides a convolutional feature map. Instead of using the selective search algorithm to identify the region proposals made in previous iterations , a separate network is used to learn and predict these regions. The predicted region proposals are then reshaped using a region of interest (ROI) pooling layer, which is then used toclassify the image within the proposed region and predict the offset values for the bounding boxes. The strategy behind the region proposal network (RPN) training is to use a binary label for each anchor, so the number one will represent the presence of an object and number zero the absence; with this strategy any IoU over 0:7 determines the object's presence and below 0:3 indicates the object's absence. Thus a multitask loss function shown in Equation (3) is minimized during the training phase.

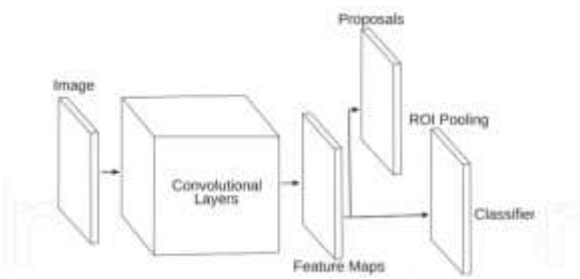

Figure 2.3: Faster RCNN acts as a single, unified network for object detection

The region proposal network module serves as the "attention" of this unified network

$$L(\{p_i\}, \{t_i\}) = 1/N_{cls} \sum L_{cls}(p_i, p_i') + \lambda\ 1/N_{reg} \sum p_i' L_{reg}(t_i, t_i')\text{……………………………………………...}(3)$$

where i is the index of the anchor in the batch, $p_i$is the predicted probability of being an object, p_i'is the ground truth probability of the anchor, t_i is the predicted bounding box coordinate, t_i'is the ground truth bounding box coordinate, and $L_{cls}$and L_{reg} are the classification and regression loss, respectively

Voice Output: Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech computer or speech synthesizer, and can be implemented in software or hardware products. A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech.

Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diphones provides the largest output range, but may lack clarity. For specific usage domains, the storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output.The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood clearly. An intelligible text-to-speech program allows people with visual impairments or reading disabilities to listen to written words on a home computer.

OpenCV: OpenCV (Open Source Computer Vision Library) is a library of programming functions mainly aimed at real-time computer vision. Originally developed by Intel, it was later supported by Willow Garage then Itseez (which was later acquired by Intel). The library is cross-platform and free for use under the open-source BSD license. Python is dynamically typed and garbage-collected.It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

III. Result

This can detect and recognize various categories of obstacles that may be faced while walking, objects of daily use and appliances, different types of vehicles, food, etc. This is beneficial for the users from different sectors like Banking, Travel and Tourism, Food and Beverages, Education, etc. Feedbacks taken from people working in banking sector prove the same point under consideration. A detection application is intended to help people with a visual impairment to find more precisely where objects are located through the proposed system.

In this paper, we design an object detection system using a deep learning object recognition technique and voice recognition technology. This system's voice synthesis provides convenience features for the visually impaired. As one of the areas where deep learning technology can be applied, our study was conducted by focusing on how to effectively aid the blind. As a result, voice recognition and voice guidance technologies were added to the system, and its performance was tested.

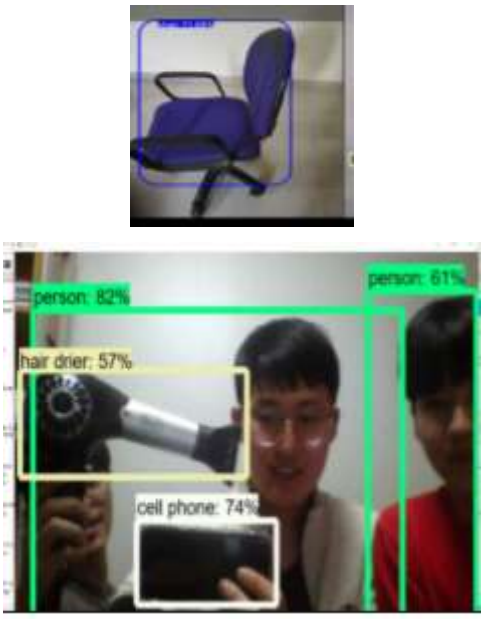

Figure 3.1: Object recognition using Python

Figure 3.2: Person detected

Figure 3.3: Person and bottle detected

Figure 3.4: Two persons detected

IV. Conclusion

This study can be used widely to provide the blind with privacy and convenience in everyday life. Also, it is expected to be applied to industrial areas where diminished visibility occurs, such as coal mines and sea beds, to greatly help production and industrial development in extreme environments. The study aims to enable people with visual impairment to live more independently. People with visual impairment will be able to overcome some threats that they may come across in their day to day life that may be either while reading a book or traveling through the city by making efficient use of the application and its associative voice feedback. Therefore, it will help to prevent possible accidents. The mobile devices can be carried easily and the camera of the device can be used to detect object from the surroundings and give output in audio format. Thus, helping visually impaired people to 'See Through the Ears'.

References:

1. Global data on visual impairment, World Health Organization, https://www.who.int/blindness/publications/globaldata/en/.

2. Tom M. Mitchell "Machine Learning". McGraw Hill Education 2017.

3. Rafael C. Gonzalez and Richard E. Woods "Digital Image Processing". Pearson 2018.

4. Aditya Raj, Manish Kannaujiya, Ajeet Bharti, Rahul Prasad, Namrata Singh, Ishan Bhardwaj " Model for Object Detection using Computer Vision and Machine Learning for Decision Making " International Journal of Computer Applications (0975 – 8887) Volume 181 – No. 43, March 2019.

5. Selman TOSUN, Enis KARAARSLAN "Real-Time Object Detection Application for Visually Impaired People: Third Eye". IEEE Conferences 2018.

6. "Jayshree R Pansare, Aditi Gaikwad, Vaishnavi Ankam, Priyanka Karne and Shikha Sharma "Real - Time Text Reader" International Journal of Computer Applications 182(34):42-45, December 2018.

8. OpenCV, https://opencv.org/ .

9. Python programming language, https://www.python.org/ .

10. Tensorflow, https://www.tensorflow.org/

11. Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi "You Only Look Once: Unified, Real-Time Object Detection". Cornell University, Jun 2015

12. Google "Google Cloud Services", https://cloud.google.com/

13. Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, Jiajun Liang "EAST: An Efficient and Accurate Scene Text Detector". Cornell University, Jul 2017.

14. Tesseract-Ocr, https://github.com/tesseract-ocr

15. Google Cloud Text-to-Speech, https://cloud.google.com/text-to-speech.