IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 6, June 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906170 |

Speech Activated Object Detection with Visual Recognition

Prajwal M[1], Raveesh M K[2], Rohit Sajjan[3], S G Prashanth Raju[4], Prashanth J[5]

UG Students, Department of Computer Science & Engineering, BNM Institute of Technology, Bengaluru, Karnataka, India. [1,2,3,4]

Professor, Department of Computer Science & Engineering, BNM Institute of Technology, Bengaluru, Karnataka, India. [5]

ABSTRACT: Object detection is a branch of computer vision that uses bounding boxes to recognize the occurrences of semantic items in an image. With the advent of image processing and neural networks, we don't need as much as new technology while the project intends to enable society to experience the world independently with the aid of a speech-based feedback system. This paper proposes methods for identifying and locating specific types of objects by navigating the users to reach that object. The goal of the project is to convert the visual world into an auditory world that can convey artefacts about their spatial locations. To give the location of the object in the camera's view, the annotated text is transformed into audio responses. The project uses artificial intelligence to characterize people, text, and objects. The project employs powerful machine learning algorithms such as MobileNet-SSD and CNNs (Convolution Neural Networks) to process images using a pre-trained model and uses Google Text to Speech, Pyttsx for speech to describe the situation in real time for the users. The experiment is conducted on datasets like Pascal VOC dataset and COCO dataset covering wide range of objects.

KEYWORDS: Object detection, Deep neural network, Open CV, Artificial Intelligence, MobileNet SSD, Google Text to Speech, Pyttsx, Pascal VOC, COCO.

I. INRODUCTION

As a major breakthrough in artificial intelligence, deep learning has achieved very impressive success in solving grand challenges in many fields including speech recognition, natural language processing, computer vision, image and video processing, and multimedia. India presently has around 12 million visually impaired individuals against 39 million all around, which makes India home to 33% of the world visually impaired populace. Although they can develop alternative approaches to deal with daily routines, they also suffer from certain navigation difficulties as well as social awkwardness. Computer Vision is a promising to use state-of-art techniques to help people with vision loss. Through this project, we want to explore the possibility of using the hearing sense to understand visual objects. The sense of sight and hearing share a striking similarity: both visual object and audio sound can be spatially localized. It is not often realized by many people that we are capable at identifying the spatial location of a sound source just by hearing it with two ears. In our idea, we build a real-time object detection and position estimation pipeline, with the goal of informing the user about surrounding object and their spatial position using binaural sound.

Paper is organized as follows. Section II describes about literature survey and prior work done in the field of object detection. Section III describes about the methodology followed and the various modules covered in the proposed system. Section IV, V describes about the system architecture. Section VI shows the results and analysis of the model which is trained on Pascal VOC and COCO dataset. Lastly Section VII and VIII discuss the future enhancements and conclusion.

II. LITERATURE SURVEY

The novel invention “Real-Time Visual Recognition with Results Converted to 3D Audio” [1] by Rui (Forest) Jiang, Qian Lin, and Shuhui Qu says that the objects detected from the scene are represented by their names and converted to speech. Their spatial locations are encoded into the 2-channel audio with the help of 3D binaural sound simulation. Video is captured with a portable camera device and is streamed to the server for real-time image recognition with existing object detection models (YOLO), the location and the size of the bounding boxes from the detection algorithm.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906170 |

3D sound generation application based on Unity game engine renders the binaural sound with locations encoded. The prototype device is tested in a situation simulating a blind people being exposed to a new environment.

J. Redmon, S. Divvala, R. Girshick and A. Farhadi presents disclosure "You Only Look Once: Unified, Real-Time Object Detection" [2] is related to a new approach to object detection. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation, base YOLO model processes images in real-time at 45 frames per second. Fast YOLO, processes an astounding 155 frames per second while still achieving double the mAP of other real-time detectors. Compared to state-of-the-art detection systems, YOLO makes more localization errors but is less likely to predict false positives on background. It out performs other detection methods, including DPM and R-CNN.

Okeke Stephen, Deepanjali Mishra, Mangal Sain presents an invention "Real Time object detection and multilingual speech synthesis" [3] where a new technique for real time deep learning based image detection with multilingual neural text-to-speech (TTS) synthesis; to generate different voices from a single model. In this work, we show improvement to the existing single lingual approach for a single-model based neural text to speech synthesis. This model, constructed with higher performance building block of a neural network (Inception4 model), demonstrated high level significant audio signal quality improvement on the images detected in real life. They show that a single deep learning model with a single neural TTS system can generate multiple languages with unique voices and display them in real life environment. They adopted transfer learning method for the image detection and recognition purpose and retrained the top layer of the model. This work introduces user friendly image-to-speech tracking system for easy navigation and valuable information extraction especially for the visually impaired people with multiple language capabilities.

Heetika Gada1 ,Vedant Gokani2 , Abhinav Kashyap3 , Amit A. Deshmukh proposed their invention "Object Recognition for The Visually Impaired" [4] with objective to recognize objects in real time and allot the objects to the classes that are previously defined. The algorithms that we utilized are more computationally efficient. Previously, object detection was done using RFID and IR technologies which required dedicated hardware. But with the advent of image processing and neural networks, we require almost no new hardware. Almost everything has camera these days from pens to mobile phones. This has given rise to a new field called computer vision i.e. using pictures and videos to detect, segregate and track objects or events so that we can "understand" a real world scenario.

III. METHODOLOGY

The model uses MobileNet-SSD algorithm for object detection and Google Speech Recognition package (gTTS) for audio processing along with Pyttsx package. Pascal VOC dataset is used for object detection; segmentation and captioning which has 20 different classes making it very versatile for object detection.

The proposed model mainly contains 3 modules.

1) **Speech to Text conversion**: This module accepts the input command from the user in order to activate the model. The input speech is recognized and interpreted for performing further actions.

2) **Object Detection**: Object Detection is a computer technology related to computer vision and image processing that is used in digital images and videos to recognize instances of symbolic objects of a certain type (such as humans, homes or cars). The model used here for object detection is MobileNet-SSD (Single-Shot Multibox Detection) algorithm that runs through a variation of an extremely complex Convolution Neural Network architecture called the MobileNet which is incorporated in our project so that it can help the user to identify commonly used objects in daily life.

3) **Text to Speech conversion**: The result of the object detection which is a textual output is converted to speech by using Google API. It recognizes the textual data placed in front of the camera module with the help of text recognition and converts it into speech, thus helping visually impaired (user) to identify the textual information in front of them. This module is the final Output of the model.

The flow diagram of proposed model is as shown in figure (Fig-1). The model waits for all the necessary libraries and weights to be instantiated and the model is loaded. The input voice command "Help Me" triggers the model and accepts the voice query from the user. The voice input is processed and the model predicts the presence of the object in the

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906170 |

frame. If the object is found, the model outputs the class name of the object predicted by labeled voice output. If the object is not found in the frame, the model again waits for further command.

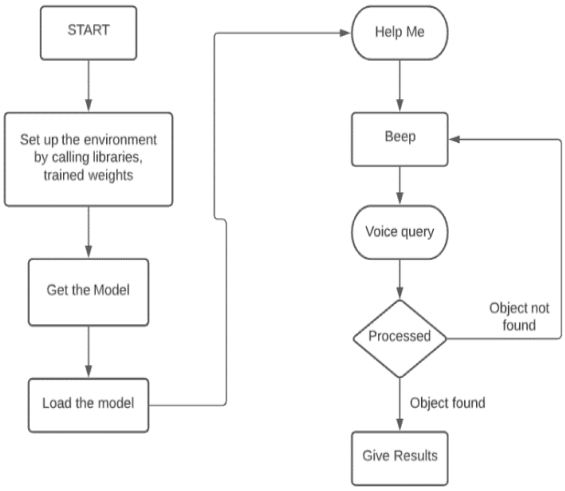

Fig-1: Flow diagram

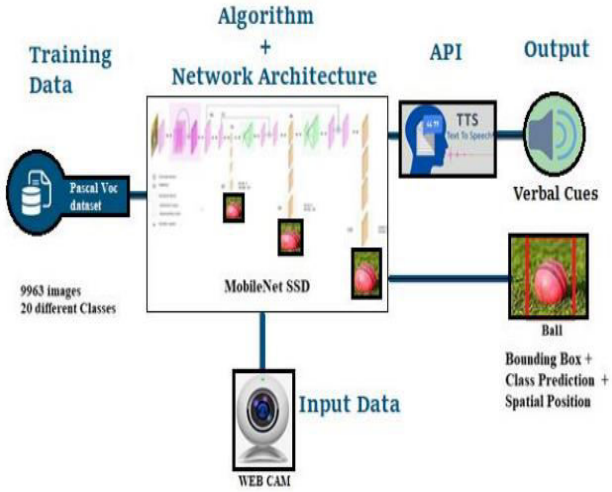

Fig-2: Proposed system

IV. PROPOSED SYSTEM

The proposed model must detect the objects and give audio feedback to visually impaired people. The most important thing in training a deep learning model is to choose a suitable dataset and object detection algorithm. There are various datasets available and the Pascal VOC dataset is the most popular one which contains various class objects with annotations. SSD object detection algorithm is one of the recent algorithms which have very good performance compared to other algorithms like RCNN and YOLO. First, we train our MobileNet SSD model using the Pascal VOC dataset. After the model is successfully trained it will now be capable of detecting objects given an input from the camera. The model detects all the objects that are present in the image and draws a bounding box around them with a label denoting the class of the object it belongs to and the spatial position of the object relative to the user. The output of the model will be in the form of text and is converted to speech using Google text to speech API. This audio will be played to the user through the user's device (headphone). The proposed system is as shown in Fig 2.

The system is designed to perform the following tasks:

- Design a system to detect objects and give audio feedback to user using Deep Learning model.
- Obtain audio commands and image from the user as input to the model.
- Detect class of objects in the image.
- Provide audio feedback of the result to the user.

V. SYSTEM DESIGN

The SSD approach is based on a feed-forward Convolutional network that produces a fixed-size collection of bounding boxes and scores for the presence of object class instances in those boxes. Single Shot MultiBox Detector is one of the many object detection algorithms. It can detect objects with over 79% mean average Precision at 59fps on COCO and VOC dataset.SSD only needs an input image and ground truth boxes for each object during training. We evaluate a small set of default boxes of different aspect ratios at each location in several feature maps with different scales (e.g. 8×8 and 4×4). For each default box, we predict both the shape offsets and the confidences for all object categories ((c1, c2, · · · , cp)). At training time, we first match these default boxes to the ground truth boxes which are treated as positives and the rest as negatives as shown in Fig-3.

SSD ARCHITECTURE: SSD has two components: a backbone model and SSD head. *Backbone* model usually is a pre-trained image classification network as a feature extractor. This is typically a network like ResNet trained on ImageNet from which the final fully connected classification layer has been removed. Thus we are left with a deep neural network that is able to extract semantic meaning from the input image while preserving the spatial structure of the image at a lower resolution. For ResNet34, the backbone results in a 256 7x7 feature maps for an input image. The *SSD*

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906170 |

head is just one or more Convolutional layers added to this backbone and the outputs are interpreted as the bounding boxes and classes of objects in the spatial location of the final layers activations. In the figure below (Fig-4), the first few layers (white boxes) are the backbone, the last few layers (blue boxes) represent the SSD head. The localization loss is the mismatch between the ground truth box and the predicted boundary box. SSD only penalizes predictions from positive matches. We want the predictions from the positive matches to get closer to the ground truth. Negative matches can be ignored. The localization loss between the predicted box l and the ground truth box g is defined as the smooth L1 loss with cx, cy as the offset to the default bounding box d of width w and height h.

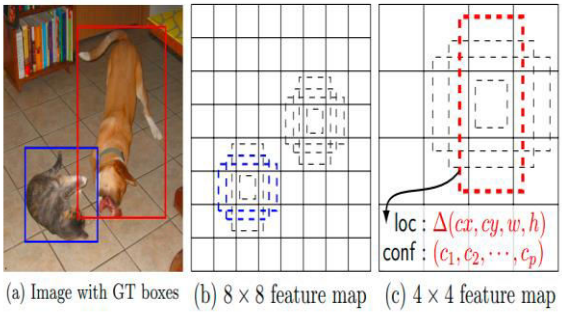

Fig-3: Training the model using SSD algorithm

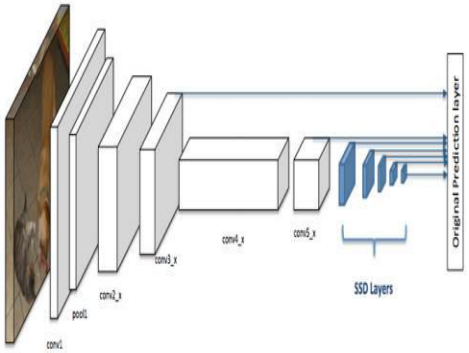

Fig-4: Architecture of SSD detector

The confidence loss is the loss of making a class prediction. For every positive match prediction, we penalize the loss according to the confidence score of the corresponding class. For negative match predictions, we penalize the loss according to the confidence score of the class "0": class "0" classifies no object is detected. It is calculated as the softmax loss over multiple classes confidence c (class score) where N is the number of matched default boxes.

The localization loss between the predicted box $\boldsymbol{l}$ and the ground truth box $\boldsymbol{g}$ is defined as the smooth L1 loss with $\boldsymbol{cx}$, $\boldsymbol{cy}$ as the offset to the default bounding box $\boldsymbol{d}$ of width $\boldsymbol{w}$ and height $\boldsymbol{h}$.

$$L_{loc}(x, l, g) = \sum_{i \in Pos}^{N} \sum_{m \in \{cx, cy, w, h\}} x_{ij}^{k} \text{smooth}_{\text{L1}}(l_i^m - \hat{g}_j^m)$$

$$\hat{g}_j^{cx} = (g_j^{cx} - d_i^{cx})/d_i^w \qquad \hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy})/d_i^h$$

$$\hat{g}_j^{w} = \log\left(\frac{g_j^w}{d_i^w}\right) \qquad \hat{g}_j^{h} = \log\left(\frac{g_j^h}{d_i^h}\right)$$

$$x_{ij}^p = \begin{cases} 1 & \text{if } IoU > 0.5 \text{ between default box } i \text{ and ground true box } j \text{ on class } p \\ 0 & \text{otherwise} \end{cases}$$

It is calculated as the softmax loss over multiple classes confidences $\boldsymbol{c}$ (class score).

$$L_{conf}(x, c) = -\sum_{i \in Pos}^{N} x_{ij}^p log(\hat{c}_i^p) - \sum_{i \in Neg} log(\hat{c}_i^0) \quad \text{where} \quad \hat{c}_i^p = \frac{\exp(c_i^p)}{\sum_p \exp(c_i^p)}$$

where $\boldsymbol{N}$ is the number of matched default boxes.

The final loss function is computed as:

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

Where, N is the number of positive matches and α is the weight for the localization loss.

OVERVIEW OF DATA FLOW: The core concept used in implementing the system is Speech to Text conversion module, Object detection module and Text to Speech module. The implementation of this concept requires the following steps:

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906170 |

1) The user gives commands to the system in the form of speech which is converted to text format that is understood by the by Object detection model using speech API.
2) The command that is converted to text triggers the model which captures multiple images from the webcam. These images are then fed into the model for object detection. The model predicts the class that the object belongs to in the dataset.
3) The prediction from the model is of two parts: bounding box which is an imaginary rectangle box that serves as a point of reference for object detection and spatial position which is the position of object relative to the user. These both are converted to speech using Pyttsx API and played to the user.

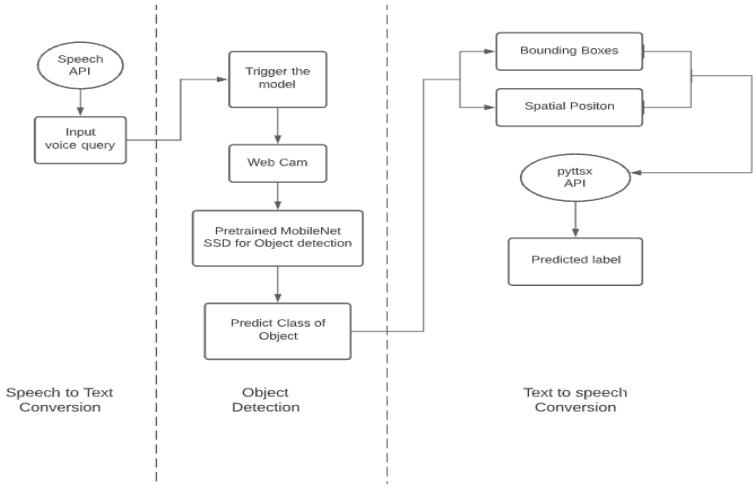

Fig-5: Overview of dataflow

VI. EXPERIMENTAL RESULTS

In order to fully test that all the requirements of an application are met, there must be at least two test cases for each requirement: one positive test and one negative test. Here in this approach, with the presence of the object in the frame, the model narrates the scene to the user as output. If the object is not found it returns "object not found" as output. The model draws a bounding box and outputs the label name of the object along with its position. The model is also able to give warning to the users if the object is very close to the frame. Following are results of some of the objects detected by the model:

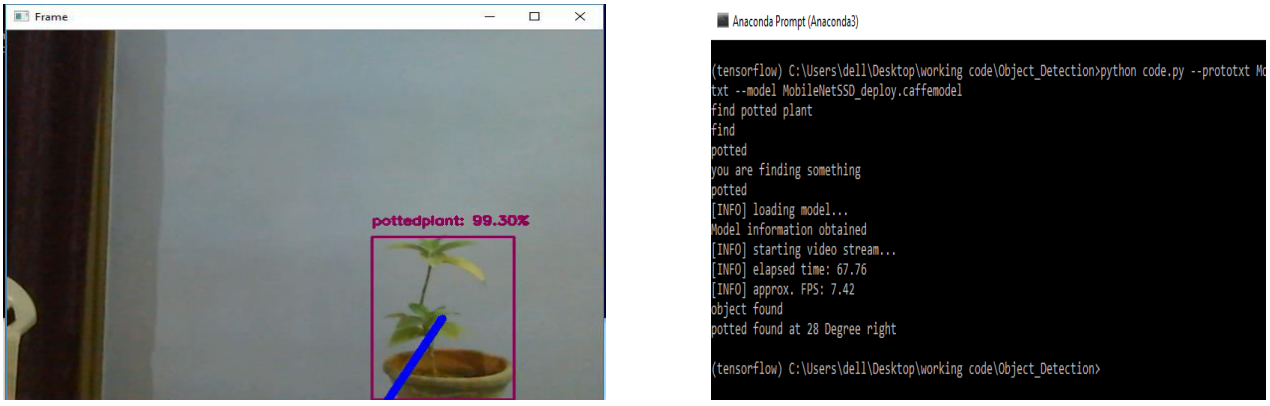

Fig-6: Working of the system which is taking potted plant as input object to find and returning its spatial position as : Object found at 28 degree right.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906170 |

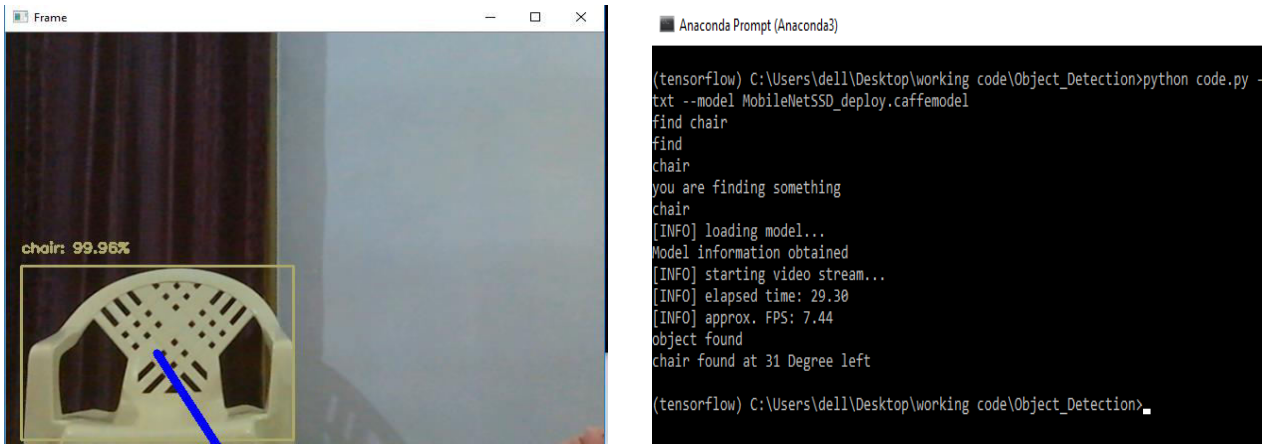

Fig-7: Working of the system which is taking chair as input object to find and returning its spatial position as: Object found at 31 degree left.

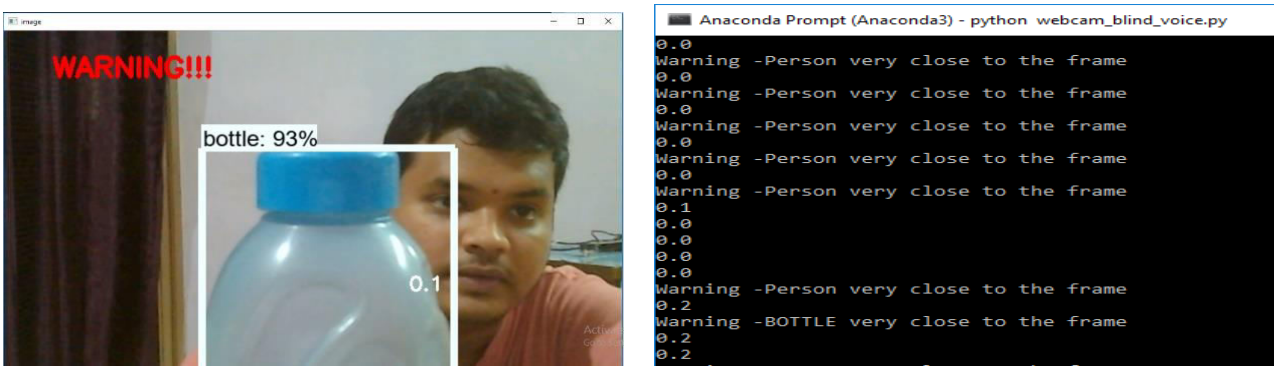

Fig-8: Working of the system which is warning the user that the objects like person and bottle are very nearer to the frame.

The following tables discuss the accuracy of Object detection for various classes of objects using SSD Algorithm.

Objects used	Accuracy
Bottle	97.30%
Car	96.20%
Chair	99.96%
Dining Table	98.41%
Dog	93.72%
Monitor	93.71%
Motor Bike	89.19%
Potted Plant	99.30%
Person	99.97%
Sofa	98.74%

Table-1 Accuracy of object detection using Pascal VOC dataset

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906170 |

Objects used	Accuracy
Bench	94.32%
Umbrella	96.75%
Shoe	98.36%
Knife	92.14%
Refrigerator	87.38%

Table-2 Accuracy of object detection using COCO dataset

VII. FUTURE SCOPE

This project is for blind people who are unable to see the vibrant and lovely world around them; our endeavour will assist them in leading a more fulfilling life. This project will allow one to understand what object is in front of him, and our team will be able to improve this product through continuous research and development by feeding more data to the Deep Learning algorithm, increasing the accuracy of the model as well as the algorithm's ability to recognize more objects. Surveillance systems, face identification, fault detection, character recognition, and other applications can all benefit from the object recognition system. The goal of this thesis is to create an object recognition system that can narrate the scene in the picture. More characteristics can be used for feature extraction and the classifier used for recognition can be improved further to determine the performance of the object recognition system. This study can be extended for new feature extraction approach for extracting global features and obtaining local features from the study area

VIII. CONCLUSION

Both Deep Learning and the Computer Vision technologies gave us the capacity to develop the model that will be of real advantage to the individual in need. The project harnesses the power of AI to describe people, text and objects. It can tell the users what is around them. The project uses advanced analytics and machine learning algorithms for image segmentation like MobileNet SSD, CNNs (Convolution Neural Networks) to process the image using a trained model and narrate the scene in the picture to the users in real-time. Once the analytics part is done the image description so created is converted into speech format and played using a speaker or earphone so that the user can hear the description about the object. Our system will be beneficial to the visually impaired, through the use of this model. This approach aims to provide a graphical user interface such that any device which includes a camera module should detect the presence of the object and reads the text that is being proposed. The project focuses on providing a system which is user friendly and cost effective with minimal hardware configuration in providing accurate results.

REFERENCES

[1] Jiang, Rui & Lin, Qian & Qu, Shuhui., “Let Blind People See: Real-Time Visual Recognition with Results Converted to 3D Audio” , Stanford University, March 2016.

[2] 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, pp. 779-788, doi: 10.1109/CVPR.2016.91.

[3] O. Stephen, D. Mishra and M. Sain, "Real Time object detection and multilingual speech synthesis," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019, pp. 1-3, doi: 10.1109/ICCCNT45670.2019.8944591.

[4] H. Gada, V. Gokani, A. Kashyap and A. A. Deshmukh, "Object Recognition for the Visually Impaired," 2019 International Conference on Nascent Technologies in Engineering (ICNTE), 2019, pp. 1-5, doi: 10.1109/ICNTE44896.2019.8946015.

[5] Kwon Lee and Chulhee Lee Yonsei, ” Fast Object Detection Based on Color Histograms and Local Binary Patterns” ,University Seoul, South Korea ,Wireless Personal Network Team Electrics and Telecommunications Research Institute Deajeon , South Korea, 2012.

[6] Samruddhi Deshpande &Ms.Revathi Shriram ,” Real Time Text Detection and Recognition on Hand Held Objects to Assist Blind People” , International Institute of Information Technology, International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT),IEEE ,2016.

[7] Leo Abraham, Nikita Sara Mathew ,” VISION- Wearable Speech Based Feedback System for the Visually Impaired using Computer Vision ” , Saintgits College of Engineering, Fourth International Conference on Trends in Electronics and Informatics (ICOEI 2020) IEEE , 2020.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906170 |

[8] KedarPotdar, Chinmay Pai and SukrutAkolkar, "A Convolutional Neural Network based Live Object Recognition System as Blind Aid", arXiv:1811.10399v1 [cs.CV] 26 Nov 2018 https://arxiv.org/pdf/1811.10399.pdf

[9] Rahul Kumar and SukadevMeher, "Assistive System for Visually Impaired using Object Recognition, M.Sc. Thesis at Department of Electronics and Communication Engineering, National Institute of Technology Rourkela, Rourkela, Odisha769 008, India, May 2015.

[10] Chucai. Yi , Y. Tian. and Areis .Arditi, "Portable Camera-BasedAssistive-Text-and-Product Label Reading From Hand-Held Objects for Blind Persons"IEEE/ASME TRANSACTIONS ON MECHATRONICS,2018.

[11] C. H. Lin ; P. H. Cheng ; S. T. Shen, "Real-time dangling objects sensing- A preliminary design of mobile headset ancillary device for visually impaired" 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, August 2014.

BIOGRAPHY

Prajwal M
Undergraduate, Dept of CSE, BNM Institute of Technology, Bengaluru, Karnataka, India.

RAVEESH M K
Undergraduate, Dept of CSE, BNM Institute of Technology, Bengaluru, Karnataka, India.

ROHIT SAJJAN
Undergraduate, Dept of CSE, BNM Institute of Technology, Bengaluru, Karnataka, India.

S G PRASHANTH RAJU
Undergraduate, Dept of CSE, BNM Institute of Technology, Bengaluru, Karnataka, India.

Prof. PRASHANTH J
Assistant Professor, Dept of CSE, BNM Institute of Technology, Bengaluru, Karnataka, India.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details