International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 05 | May 2021 www.irjet.net p-ISSN: 2395-0072

Image based Voice Assistance for Visually Impaired

Madhura Shirodkar[1], Shubham Nahar[2], Sohan Ranadive[3], Shaktivel Thevar[4], Shubham Tippe[5]

[1]*Professor, Department of Electronics and Telecommunications, Xavier Institute of Engineering, Mumbai, India*
[2,3,4,5]*Student, Department of Electronics and Telecommunications, Xavier Institute of Engineering, Mumbai, India*

Abstract - *Nowadays Most of the blind rely on the conventional white cane. It's capability to deliver navigational freedom is limited. In our project the system will tell the user about the scene in front of them by capturing an image. The captured image will be processed and the objects and text in the image will be given as the output to the user in real time through voice. In this project we use the Raspberry Pi to process the data and Raspberry Pi Camera to detect the environment using Machine Learning and Deep Learning. We also add one switch in our project for the end user to decide when they want to explore the environment by clicking images. Our project Is divided in to two parts one is object detection and other is text detection. For object detection we used YOLO algorithm. Using yolo detection is a simple regression problem which takes an input image and learns the class probabilities and bounding box coordinates and predicts the classification score for each box for every class in training. For text detection we used SHTR technique using TensorFlow to detect the text within images and using pyttsx3 we can convert the text messages in to speech format for the users using earphone.*

Key Words: **Electronic travel aids(ETA), Raspberry Pi, YOLO object detection algorithm, Optical Character Recognition (OCR)**

1.INTRODUCTION

The World Health Organization (WHO) Fact reported that there are 285 million visually-impaired people worldwide. Among these individuals, there are 39 million who are blind in the world . More than 1.3 million are completely blind. Over the past years, blindness that is caused by diseases has decreased due to the success of public health actions. However, the number of blind people that are over 60 years old is increasing by 2 million per decade. Unfortunately, all these numbers are estimated to be doubled by 2020. This issue stands alone as a great challenge for the scientific community to develop a system able to assist visually impaired people to obtain verbal descriptions of the surrounding environment and increase their independence. The simplest and the most affordable navigations and available tools are trained dogs and the white cane. The white cane is exceptionally restricted in its capability to deliver navigational freedom to its clients. Furthermore, the visually impaired people still require the help of sighted people to lead them towards the destinations. These deficiencies drive the requirement for research on evolving inventive navigational frameworks for the visually impaired people.

Many innovative explanations usually known as electronic travel aids (ETA) have hence been proposed and executed; yet none have been broadly effective in enhancing the flexibility and lives of the blinds due to the disadvantages like limited features, one way communication and bulkiness due to use of many sensors.

2. LITERATURE SURVEY

2.1 CNN Based on TensorFlow

In CNN the network is consists of an input layer, a convolutional layer, a pooling layer, a fully connected layer, and an output layer. It is mainly divided into two processes: forward and back propagation. The former finally outputs the prediction result through the network structure, and the latter performs parameter adjustment according to the difference between the prediction result and the actual value. The input layer is image data that needs to be input, and it is generally a matrix of three dimension. The convolutional layer is a fixed size filter convoluted with the image of the previous layer to extract eigenvalues in the image. In convolution convolve the input image with the weight vector and the bias value then applied to the relu activation function. In relu functions those pixel values has chance to be grey Rounded a perfect grey and those chance to be bright rounded perfect bright. Finally maximize the pooling of the output. In pooling the image size get reduce. Similarly In this paper initialize the weight and bias value of the second convolutional layer and another pooling function after getting the output from second convolution layer. We can create model using only one convolution also but if we use two-time convolution then we get output more accurate than one time convolution.

2.2 Image-based Sequence Recognition

In CRNN model, the component of convolutional layers is made by taking the convolutional and max-pooling layers from a standard CNN model

and fully-connected layers are removed. A deep bidirectional Recurrent Neural Network is built on the top of the convolutional layers. RNN can back-propagates the convolutional layer, allowing us to jointly train the recurrent layer and convolutional as one network. Transcription is the process to find the output of the sequence based on the highest probability of the per-frame predictions made by RNN. There exist two modes of transcription i.e., the lexicon-free and lexicon-based transcriptions. Finally, the network is trained with stochastic gradient descent (SGD). The average testing time is 0.16s/sample, as measured on IC03 without a lexicon. The approximate lexicon search is applied to the 50k lexicon of IC03, with the parameter δ set to 3. Testing each sample takes 0.53s on average. Model has 8.3 million parameters, taking only 33MB RAM (using 4-bytes single-precision float for each parameter), thus it can be easily ported to mobile devices.

2.3 Content-Based Image Retrieval Method

A machine learning algorithm is commonly used to acquire content analyzation. The need to import a large number of training images and use of many CPU hours are the two primary difficulties of using existing ML algorithms. In this research paper they used Google Cloud Vision application programming interface (API) to overcome these two shortcomings. Cloud Vision API is trained by Google; therefore, it saves computational time in obtaining image labels. In this paper they are coded in the R language, which calls Cloud Vision API. In this paper they fetch the input from their source and sent towards the Google Cloud using API (Application programming interphase) key. The google cloud haves a well-trained data set of their server. Server use their forwarded images as a testing dataset and test the images at their server using well trained datasets of google cloud. Cloud computing technique is highly powerful as compared to writing the algorithm in computer having slow processor. Normal computers take high power consumption to run a deep learning task and also gives the delay output.

2.4 Convolutional Character Networks

CharNet is the first one-stage model for E2E text recognition, which is different from the two-stage model, and it consist of two branches:(1) a character branch designed for direct character detection and recognition, and (2) a text detection branch predicting a bounding box for each text instance in an image

This one-stage model only employs a light-weight CNN-based character branch with 1.19M parameters. On the synthetic data, learning rate is set at 0.0002, while in the real-world dataset it is fixed at 0.002. 4-step iterative character detection with CharNet. Initially when CharNet is only trained on synthetic data, only 64.95% is “correct” from real-world training images. But this number increases immediately from 64.95% to 88.94% at the next step. Finally, after implementing 4 iterative steps this method is able to collect 92.65% correct words from real-world images. Whereas when calculating for End-2-End Recognition the efficiency is quite low (around 69%) although it is much higher than the existing system.

2.5 Steps Involved in Text Recognition

This paper focuses on the technology of Optical Character Recognition (OCR) and the technology of Text to Speech Synthesis (TTS). OCR helps the user to recognize text from the image and TTS enables the user to read text from image, which is synthesized into human voice and played with the help of a speaker. The main objective of the proposed system in this paper is to assist visually challenged people to read text from challenging background and pattern. The system has a Raspberry Pi connected with a camera, which captures the image of printed text. Then the captured image is processed with the help of different image processing techniques such as Skew correction, Linearization, Segmentation etc. This processed image is given to PyTesseract OCR engine for text extraction and the detected text is synthesized into human voice and played through audio system.

2.6 Optical Character Recognition Using Raspberry Pi

The computer vision and digital image processing are fast growing fields that are essential in many aspects of other areas like multimedia, artificial intelligence, robotics and much more. The paper is a study on recognizing text from an image by using different image enhancement techniques. Pre-processing, Post-processing, Feature Extraction etc are different steps involved in

recognition process. Here, several image enhancement techniques are discussed such as Spatial Image Filtering, Global thresholding, Histogram equalization, Averaging filter, Local thresholding, Segmentation etc. for text recognition. From this study it is found that better results can be obtained with less computation time and multilingual character segmentation and recognition also possible with better rate.

2.7 Intelligent assistant for visually impaired people

The paper presents an electronic travel aid (ETA) system for visually impaired people. ETAs transform information about the environment and convey that information thorough any sensory mode (say auditory i.e speech) other than visual forms. The intelligent assistance is called Tyflos (Greek word for blind) and its primary goal is to capture the data from various sensors and help the user to navigate through a 3-D dynamic environment. It is a secondary mobility aid which means it is complementary to the user's primary mobility aid (say white cane) so it helps the user in situations in which the primary aid does help much (say object detection and collision free navigation). This system is a real-time, wearable, portable, reliable and simple to use.

2.8 YOLO v3-Tiny

This paper presents the basic overview of object detection algorithms. There are basically two types of object detection algorithms-1) Two stage detectors e.g. RCNN, Fast RCNN and Faster RCNN these algorithms use region proposal networks (RPN) to generate regions of interests in the first stage and send the region proposals down the pipeline for object classification and bounding-box regression. Such models reach the highest accuracy but are typically slower. 2) One stage detector e.g. YOLO v1, v2, v3 and SSD that treat object detection as a simple regression problem by taking an input image and learning the class probabilities and bounding box coordinates. Such models have lower accuracy rates, but are much faster than two stage detectors. The paper then explains the theory about YOLO and its different versions especially YOLO v3-Tiny.

3. PROPOSED SYSTEM

This project aims to develop an electronic travel aid system for visually impaired people to help them navigate in their day-to-day life, by capturing the scene using a camera and processing it using different algorithms of neural network in Raspberry Pi to detect objects, text and give the output in the form of voice. This aid is a real time, wearable, portable, reliable and simple to use.

Image Acquisition: Images will be captured with the help a good quality head mount camera. Acquired images will be pre-processed before feeding to the next block.

Object Detection and Identification: According to the switch input given by the user, object identification will be carried out using image processing algorithms and neural network algorithms like YOLO is super-fast and can be run real time. YOLO sees the complete image at once as opposed to looking at only a generated region proposal in the previous methods like Faster-RCNN.

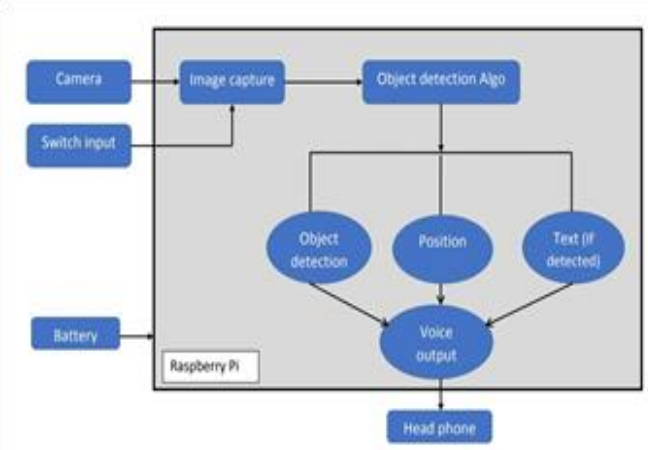

Fig -1: Block Diagram

Text Extraction from Image: Text extraction will be done if text is found on the extracted object. Text detection is the process of localizing where an image text is. Text detection as a specialized form of object detection. Our goal is to (1) detect and compute the bounding box of all Text in an image and (2) to recognize the text in the image.

1. **Position and Distance Measurement:** Position of the identified object in the image will be found like left, right or center.

   First the image will be divided into 3 parts based on the horizontal x coordinate.

   So accordingly, the image will be divided into left, center and right half.

With the help of the bounding boxes made from object detection process the coordinates of the object will be determined. Once the coordinates are known it can be decided in which half does the object lies.

2. **Voice Output:** The acquired information will be translated into voice which will be fed to headphones. Text-to-speech (TTS) is the generation of synthesized speech from text. Our goal is to make synthesized speech as natural and pleasant to listen for the visually impaired person for this purpose we are implementing pytts (python text to speech) module.

4. CONCLUSIONS

We have successfully implemented object detection using a Pre-trained data set. We will be working on making our own data set. We have also implemented text recognition and detection for image using SimpleHTR. By modifying the data set used, we can achieve more accuracy on recognizing text.

REFERENCES

[1] Liang yu et al," research and implementation of cnn based on tensorflow", iop conf. Ser.: mater. Sci. Eng. 490 042022,2019. (refer section 2.1)

[2] Baoguang Shi, Xiang Bai and Cong Yao," An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition" School of Electronic Information and Communications Huazhong University of Science and Technology, Wuhan, China,21 June 2015. (refer section 2.2)

[3] Shih-Hsin Chena, Yi-Hui Chen," A New Content-Based Image Retrieval Method Based on the Google Cloud Vision API", Conference Paper, February 2017. (refer section 2.3)

[4] Linjie Xing, Zhi Tian3, Weilin Huang, and Matthew R. Scott, "Convolutional Character Networks". (refer section 2.4)

[5] K.Karthick, K.B.Ravindrakumar, R.Francis, S.Ilankannan, 'Steps Involved in Text Recognition and RecentResearch in OCR', International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1, May 2019. (refer section 2.5)

[6] S.Thiyagarajan,Dr.G.Saravana Kumar,E.Praveen Kumar,G.Sakana,'Implementation of Optical Character Recognition Using Raspberry Pi for Visually Challenged Person', International Journal of Engineering & Technology,(2018). (refer section 2.6)

[7] N.G.Bourbakis and D.Kavraki,"AN INTELLIGENT ASSISTANT FOR NAVIGATION OF VISUALLY IMPAIRED PEOPLE", WSU, ITRI, Image-Video-Machine Vision Research Lab, Dayton Ohs. (refer section 2.7)

[8] Pranav Adarsh, Pratibha Rathi and Manoj Kumar," YOLO v3-Tiny: Object Detection and Recognition using one stage improved model", International Conference on Advanced Computing & Communication Systems (ICACCS). (refer section 2.8)