TRINETRA
A Smart Wearable Device for the Visually Impaired People

A Report Submitted
in Partial Fulfillment of the Requirements for the Course

B.Tech. Project (CS299)

by

Vaibhav Mishra
(B17CS057)

Muzzafer Ali
(B17CS037)

under the guidance of

Dr. Chiranjoy Chattopadhyay

to the

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY JODHPUR

CERTIFICATE

*This is to certify that the work contained in this thesis entitled "**TRINETRA A Smart Wearable Device for the Visually Impaired People**" is a bonafide work of **Muzzafer Ali** (**Roll No. B17CS037**) and **Vaibhav Mishra** (**Roll No. B17CS057**), carried out in the Department of Computer Science and Engineering, Indian Institute of Technology Jodhpur under my supervision and that it has not been submitted elsewhere for a degree.*

Supervisor: **Dr. Chiranjoy Chattopadhyay**

Assistant/Associate Professor,

April '19, Department of Computer Science & Engineering,

Jodhpur. Indian Institute of Technology Jodhpur, Rajasthan.

Acknowledgements

1. Dr. Chiranjoy Chattopadhyay, Assistant Professor, Department of Computer Science and Engineering, Indian Institute of Technology Jodhpur.

2. Dr. B.Ravindra, Assistant Professor, Department of Mechanical Engineering , Indian Institute of Technology Jodhpur.

Abstract

With the wonderful opportunity given to us through B.Tech project ,we focused on working on problem which inculcates values and knowledge to us ,and also fulfills the requirement for getting applied in real applications and gain some usefulness in the society. Mobility for visually impaired people in an outdoor environment has always been a challenge. There has been numerous solutions and devices to assist the blind but none of them is a complete solution and lacks in some features or other. With the current pace of increasing intelligence among machines, this is the best time to apply this intelligence to solve human problems, hence we decided to create a device fitted with a camera to understand the environment with the power of computer vision and subsequently deliver the message to the person. Various computer vision algorithms and Deep Learning techniques has been added to make it more functional and robust. Video calling service has been provided to contact in emergency situation. In this report, we explain our work done throughout the semester to create the proposed solution and give a detailed explanation on the choice of algorithm and implementation. Finally, we explain our challenges and final outcome of our work.

Contents

Chapter 1

Introduction

1.0.1 Problem choice:

Blindness is very prevalent in the society with a global estimation of approximately 1.3 billion people live with some form of distance or near vision impairment. With regards to distance vision, 188.5 million have mild vision impairment, 217 million have moderate to severe vision impairment, and 36 million people are blind. With regards to near vision, 826 million people live with a near vision impairment . This statistics was the motivation for us to work on very common but highly challenging problem which is to create a device for assisting the visually impaired people to make their life easier.

1.0.2 Introduction to Computer Vision

Computer vision is an interdisciplinary scientific field that deals with how computers can be made to gain high-level understanding from digital images or videos. From the perspective of engineering, it seeks to automate tasks that the human visual system can do.

1.0.3 Introduction to Microsoft Kinect Camera

Kinect for Xbox 360 is a combination of Microsoft built software and hardware.The device features an "RGB camera, depth sensor and microphone array running proprietary soft-

ware", which provide full-body 3D motion capture, facial recognition and voice recognition capabilities. Depth Camera works on the principle of Infrared Rays absorption and reflection .it contains an Infrared emitter and absorber which creates an 3D map bases on the captured IR waves captures by the absorber.

1.0.4 Introduction to APIs

An application programming interface (API) is a set of subroutine definitions, communication protocols, and tools for building software. In general terms, it is a set of clearly defined methods of communication among various components.

1.0.5 Introduction to Egocentric Vision

Egocentric vision or first-person vision is a sub-field of computer vision that entails analyzing images and videos captured by a wearable camera, which is typically worn on the head or on the chest and naturally approximates the visual field of the camera wearer.Consequently, visual data capture the part of the scene on which the user focuses to carry out the task at hand and offer a valuable perspective to understand the user's activities and their context in a naturalistic setting.

1.1 Functionality description

1.1.1 Real Time Object Detection

RGB video feed is taken through Kinect Camera and Object Detection Technique is applied to classify the objects present in the surrounding. We have used MobileNet running on SSD algorithm for object detection trained on COCO Dataset.

1.1.2 Depth Estimation

Kinect Camera has an IR Camera which sends the IR rays and an absorber absorbs the IR rays reflected by the objects present nearby. Hence, we get a Depth Map which gives estimation of the objects present nearby.

1.1.3 Real Time Face Recognition and Updation

One shot learning Face Recognition Technique is used to detect and recognize the faces . This helps as it requires only a small dataset of faces for training and also updation of face of new person can be done without retraining the complete model.

1.1.4 Scene Description and Image captioning

Image captioning requires the combination of both Computer Vision and Natural Language Processing to identify the objects present in the image and also to to create textual information out of it.

1.1.5 Optical Character Recognition

Optical Character Recognition is based on CNN+RNN as it contains sequential data. LSTM is the advanced version on RCNN for extracting sequential data from images. Pytessaract library based on LSTM is used for OCR.

1.1.6 Emergency Video Calling Service

A simple flask web server is launched which hosts the webcam video by sending image packets on a HTML page which can be accessed on a localhost.

1.1.7 Speech recognition and Voice enabled Assistant

Natural Language Processing is applied to recognize the human speech and convert it to text. We have used Speech Recogntion Library for NLP task. pyttsx3 library is used for

converting text to voice for guiding the Visually Impaired through the voice.

Chapter 2

Real Time Object Detetcion

2.1 System overview

In this feature we get live video from kinect Camera and then for each frame we pass it through our object detection model. Our model detect different type of objects and return its bounding box and then we show the objects with its bounding box in each frame.

Figure 2.1 shows output frame of object detection that comes from our model. We have combined the MobileNet architecture and the Single Shot Detector (SSD) framework, we arrive at a fast, efficient deep learning-based method to object detection. The model is trained on the COCO dataset therefore, it can detect 21 different type of objects.

2.2 Detailed description

2.2.1 SSDs(Single Shot Detector): Method to detect objects

We are using SSDs method for our object detection, originally developed by Google, are a balance between the Faster R-CNN and YOLO.Basically it is faster than Faster R-CNN and more accurate than YOLO and it is more compatibale with our device for object detection in real time video(order of 24 FPS).

2.2.2 MobileNets: Efficient (deep) neural networks

When building object detection networks we normally use an existing network architecture, such as VGG or ResNet, and then use it inside the object detection pipeline. The problem is that these network architectures can be very large in the order of 200-500MB. Network architectures such as these are unsuitable for resource constrained devices due to their sheer size and resulting number of computations. Instead, we can use MobileNets, another paper by Google researchers. MobileNets differ from traditional CNNs through the usage of depthwise separable convolution.

2.2.3 MobileNet SSD: Combination of SSDs and MobileNet

The MobileNet SSD was trained on the COCO dataset (Common Objects in Context) and then fine-tuned on PASCAL VOC.We can therefore detect 20 objects in images (+1 for the background class), including airplanes, bicycles, birds, boats, bottles, buses, cars, cats, chairs, cows, dining tables, dogs, horses, motorbikes, people, potted plants, sheep, sofas, trains, and tv monitors

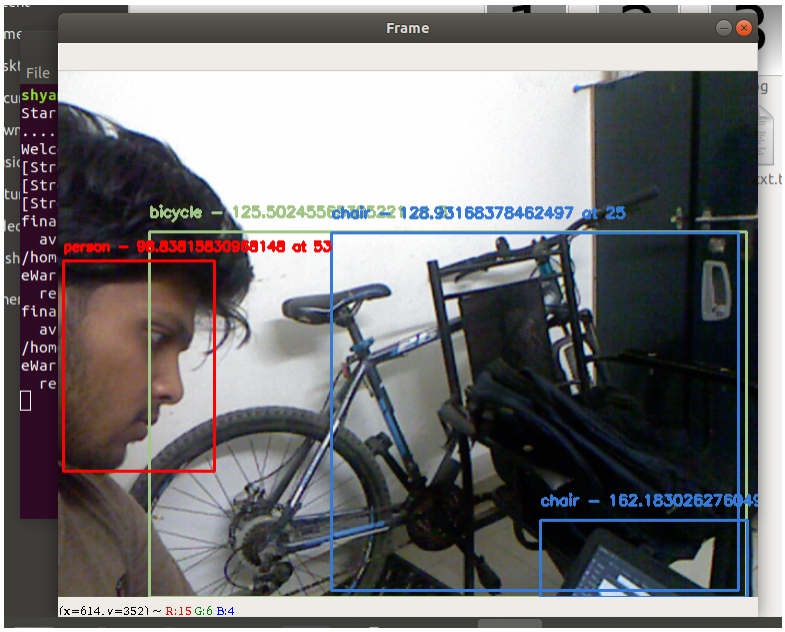

Fig. 2.1: Real Time Object Detetcion

Chapter 3

Depth Estimation

3.1 System overview

Depth Estimation is a crutial task for guiding the blind so that he does not collide with the objects present nearby him. There are various sensors available for measuring the distance for example ultrasonic sensor, laser etc. The shortcomings of such sensors is that they work only on a line of sight whereas Kinect Depth camera creates depthmap of the whole environment. We have used Microsoft Kinect XBOX which is best suitable for working on depth estimation and Computer Vision and NLP related work.

3.2 Detailed description

3.2.1 Working of Kinect Camera Sensor

Kinect Camera contains a RGB Camera , Depth Camera and an array of microphone sensors. Depth Camera present in Kinect makes it very useful for 3D applications. The Depth Camera consists of an IR rays emitter and an absorber . the IR rays are send from the emitter and the IR rays after reflection are absorbed by the absorber and a depthmap is created by the device.

Depth estimation + object detection

For obstacle detection purpose, above mentioned object detection technique is used to identify the objects present in the vicinity , then the bounding box is found and the depth pixels are extracted from the bounding box and and the mean of pixels is found over 5 frames for more accurate reading. the distance calculated from the intensity is then used to warn the person of nearby obstacles. if the object crosses a threshold of 0.5 meter then the warning sign is raised which tells the person about the obstacle and its coordinates.

Better approach for Depth Estimation

Depth Estimation done with Kinect Camera is very vague and inaccurate as the IR Camera is not very powerful, also IR rays is affected by the colour of object and its absorbtion properties and other conditions such as moisture content, presence of other heat source etc. this leads to problems in guiding the person of the accurate coordinates of the obstacle. Other advanced technique such as StereoVision can be implemented for much accurate distance calculation. Future Version of the Device may include such technique of object distance calculation.

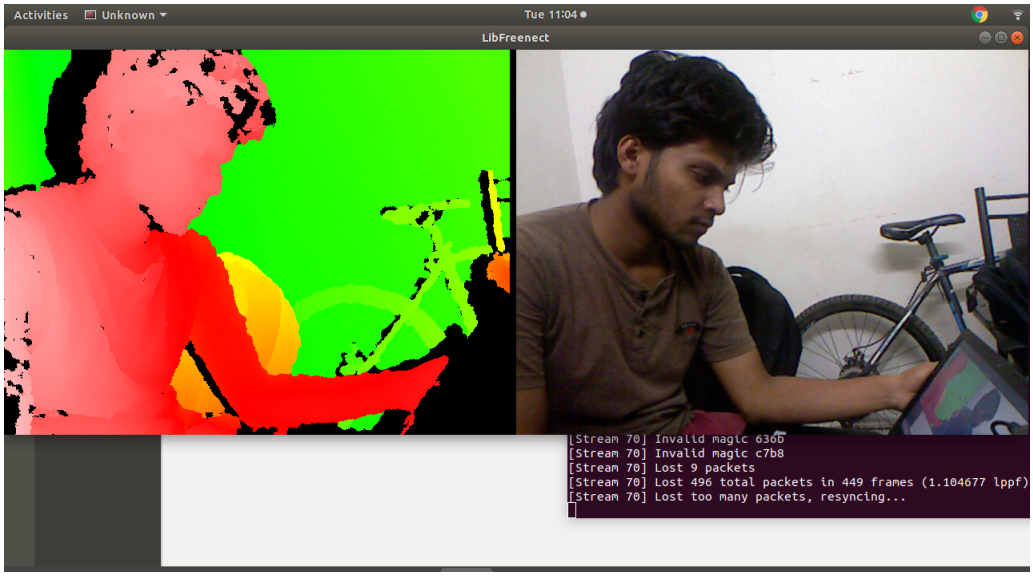

Fig. 3.1: Depth from kinect

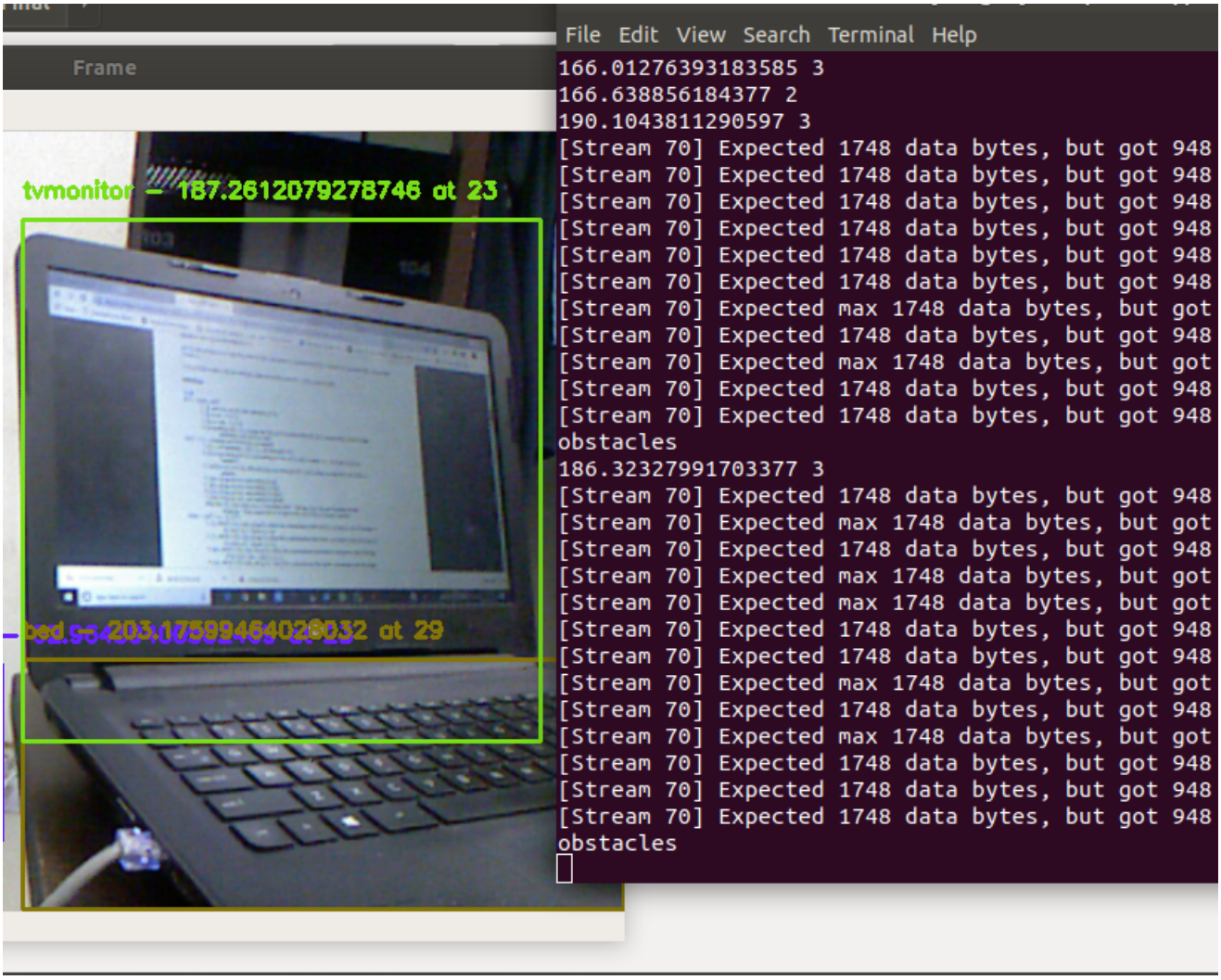

Fig. 3.2: For object at less 0.5 meter its showing obstacle

Chapter 4

Real Time Face Recognition and Updation

4.1 System overview

Before we can recognize faces in real time videos,we first need to quantify the faces in our training set. So first we have stored some faces in dataset images and for that we are using face detection method - HOG. Then we construct 128-d embedding(quantification) for each images. After creating embeddings of dataset and saving it to a file, we take each frame and first we detect all the faces, than create embeddings and than we use K-NN model for classification. Finally showing person names with its bounding box to each frame, an example is shown in Figure 4.1. And the best part of this feature is when unknown person is detected then its gives the option for training him in our dataset.

4.2 Description

4.2.1 HOG(Histogram Of Oriented Gradient)

The technique counts occurrences of gradient orientation in localized portions of an image. This method is similar to that of edge orientation histograms, scale-invariant feature trans-

form descriptors, and shape contexts, but differs in that it is computed on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for improved accuracy.We are using this method over CNN because it can detect face faster than CNN, and for real time detection we need faster method.

4.2.2 128-d embeddings

Our network quantifies the faces, constructing the 128-d embedding (quantification) for each.The general idea is that well tweak the weights of our neural network so that the 128-d measurements of the same person will be closer to each other and farther from the measurements of other persons Our network architecture for face recognition is based on ResNet-34 from the Deep Residual Learning for Image Recognition paper by He et al., but with fewer layers and the number of filters reduced by half and the network was trained over 3-million images.

4.2.3 K-Nearest Neighbors algorithm (K-NN)

K-NN is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point. So here after getting embeddings of faces from each frame we use K-NN to classify in which category it lies.

4.2.4 Live Updation

In our face recognition model when unknown person is detected then its gives option to add and train the unknown person in our data set so that next time it automatically recognize that person with taking input of its name. Basically, we load the embeddings of dataset from file that store embeddings in bytes in our program and then we add embeddings of unkown person in our data from 10 different frames and again load it to our embeddings file.

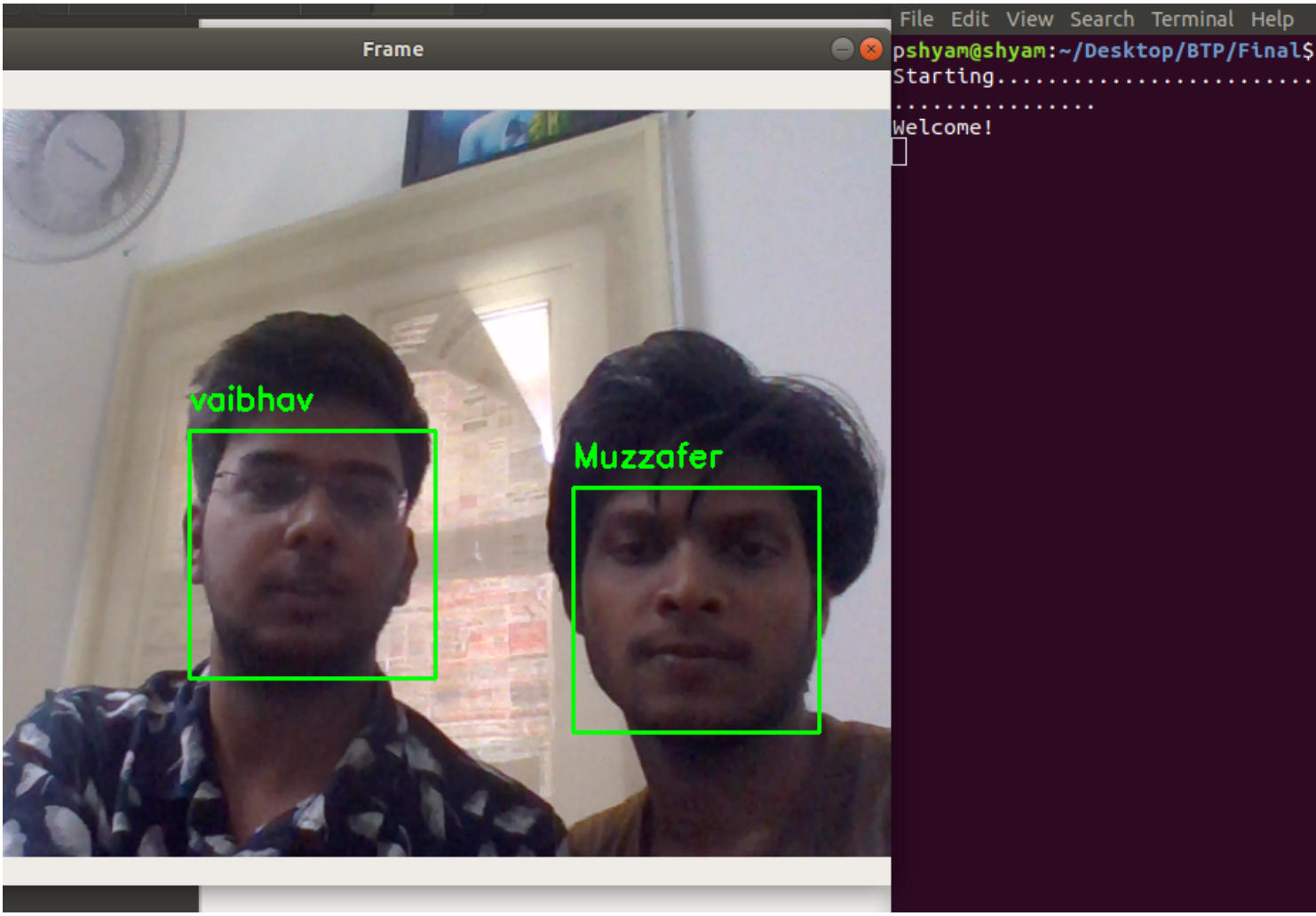

Fig. 4.1: Real Time Face Recognition

Chapter 5

Scene description and Image Caption

5.1 System overview

Scene Description is the most important feature in perspective of the visually impaired person. This feature allows to describe an image in the form of certain sentences focusing on the description of various objects and their behaviour. This is the most helpful and interactive feature out of all the present feature.

5.2 Detailed description

5.2.1 Working of Image captioning

Image captioning is done with the help of deep learning model. The task of image captioning can be divided into two modules logically one is an image based model which extracts the features and nuances out of our image, and the other is a language based model which translates the features and objects given by our image based model to a natural sentence.

The image based model uses a Convolutional Neural Network to understand various objects and other behavioural details and encodes this information which is later decoded by the language based model. Hence, CNN acts as a encoder of various information and saves the encoded information. the original paper on image captioning uses a ResNet model

for encoder. The language based model is based on RNN or LSTM which is advanced version of RNN. a dictionary of words is fed to this model and the network decodes the encoded information of the image and assigns the words to the image and create sentences out of it.

5.2.2 Our implementation of Image Captioning

Since Image Captioning requires a lot of images and a seperate vocabulary list of words to be trained, it is difficult to train the model from scratch with limited resources. For our Device, we use an API from DeepAI which takes an image as input and returns the description of image in form of json object. Later, the json object is parsed to extract the sentences which has a probability greater than 0.85 .

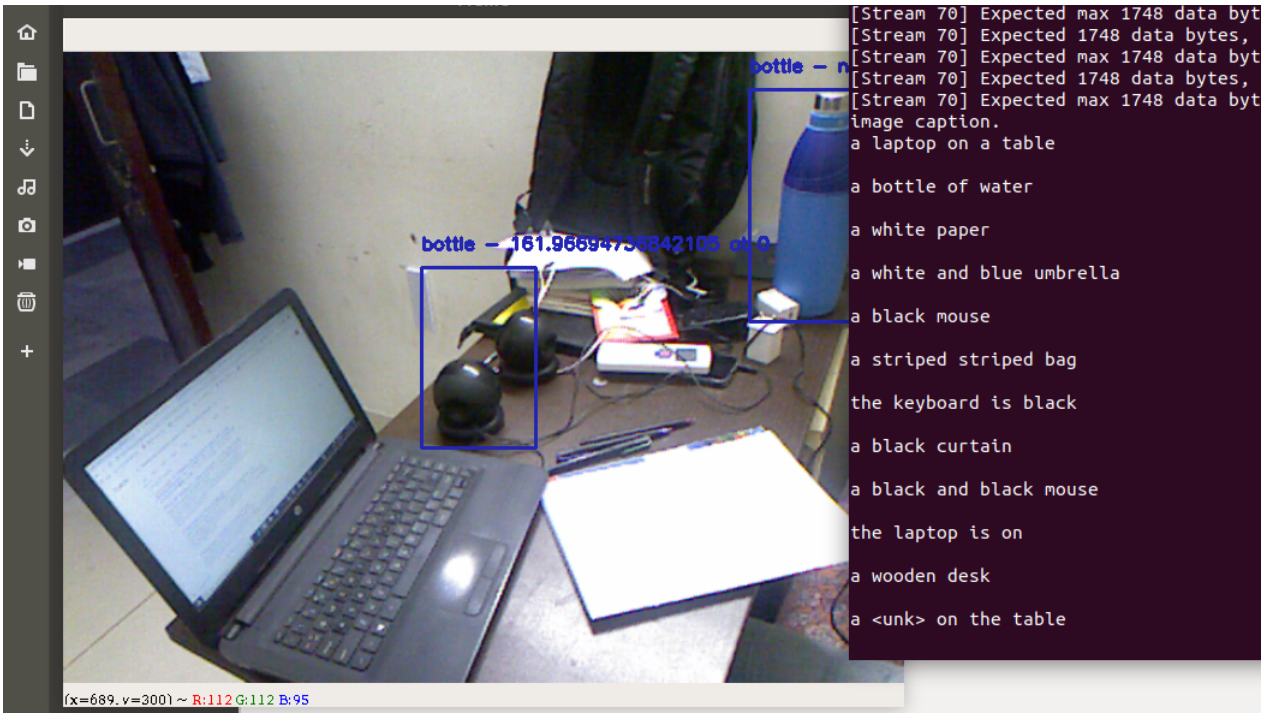

Fig. 5.1: Scene description and Image Caption

Chapter 6

Optical Character Recognition

6.0.1 System overview

OCR can help the person in recognizing signboards, identifying notes, reading books etc. Optical Character Recognition is based on CNN+RNN as it contains sequential data. LSTM is the advanced version on RCNN for extracting sequential data from images. Pytessaract library based on LSTM is used for OCR.

6.1 Detailed description

6.1.1 Working of Optical Character Recognition

The aim of OCR is to recognize not only characters but words and sentences.Hence, by simply training on characters is not going to help in this situation as this will lead to very poor accuracy and bad results. There is a sequential information linked between different characters and words in a sentence. Thus, RNN works better than simple CNN as it is used for extracting sequential information from data. The idea goes as follows: the first level is a standard fully convolutional network. The last layer of the net is defined as feature layer, and divided into feature columns. Afterwards, the feature columns are fed into a deep-bidirectional LSTM which outputs a sequence, and is intended for finding relations

between the characters. Finally, the third part is a transcription layer. Its goal is to take the messy character sequence, in which some characters are redundant and others are blank, and use probabilistic method to unify and make sense out of it.

6.1.2 Our implementation of OCR

We have used pytesseract library for OCR implementation as it uses state of the art technique for character recognition by using LSTM network. LSTM is a type of RNN which improves upon the shortcomings of RNN. Tesseract was developed as a proprietary software by Hewlett Packard Labs. Since 2006 it has been actively developed by Google and many open source contributors.

Tesseract library is shipped with a handy command line tool called tesseract. We have used this tool to perform OCR on images and the output is stored in a text file. We integrated Tesseract in Python code by using Tesseracts API.

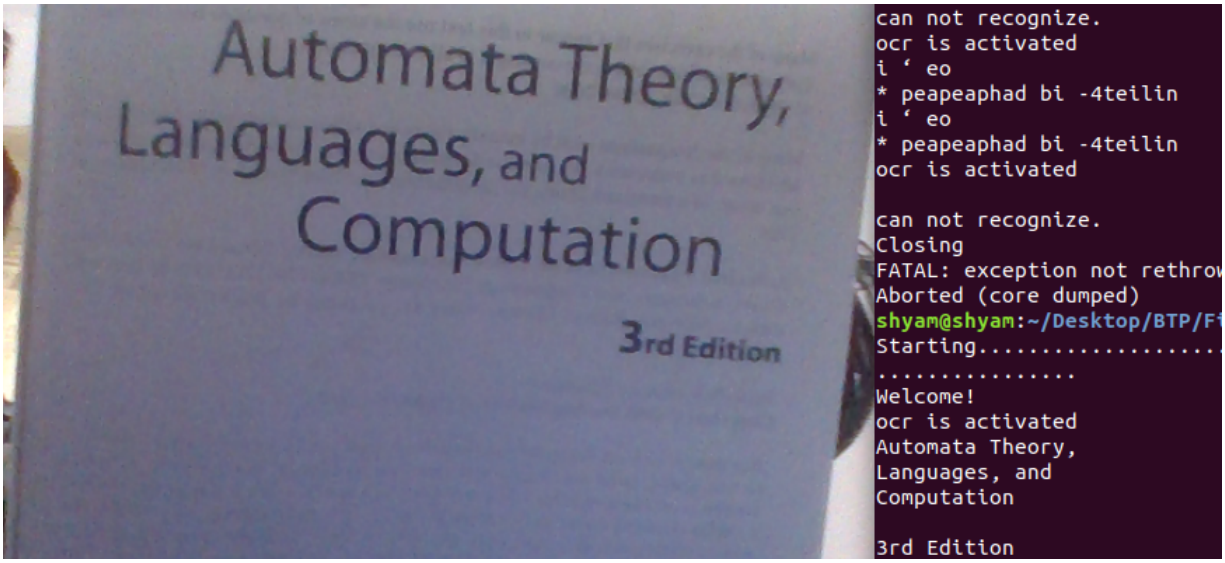

Fig. 6.1: Reading mode

Chapter 7

Emergency Video Calling Service

7.1 System overview

Imagine a situation in which the Visually Impaired Person is stuck and he could not decide upon what to do.The video calling service serves the purpose of broadcasting the video of the device to multiple user across the internet who are willing to help him. The person on the other side can see the Visually impaired person's environment and guide him to safe place.

7.2 Detailed description

7.2.1 Working of the video calling service

A simple flask web server is launched which start a HTML webpage on a localhost. later a video tag is embedded on the HTML page to show the Video stream. The video frames are extracted from the webcam and sent to other devices through socket and the extracted video frames are displayed on a HTML webpage .

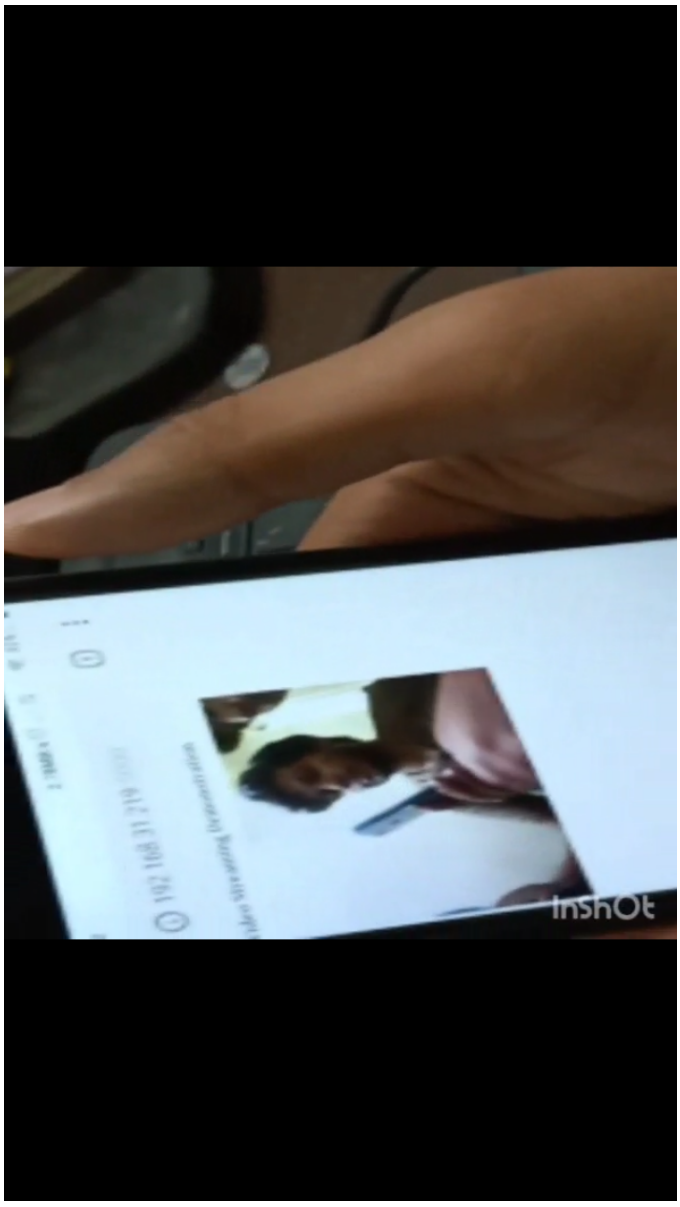

Fig. 7.1: Calling mode

Chapter 8

Speech recognition and Voice Assistant

8.1 System overview

Its big challenge for us to make our device compatible with blind people because there is no use of screen, we can't provide information to user through it. So the best option is to use audio signal to make our device interactive with user. We are providing two features voice assistant and speech recognition For voice assistant we are using the python library pyttsx3 that is offline and we will giving all information in audio form through pyttsx3 as like "Muzzafer has detected at angle 30 degree to the right", book reading, etc. For Speech recognition we are using the `speech_recognition` and pygame library of python. The library `speech_recognition` support many speech recognition engines and APIs, and for our project we are using google speech recognition API to recognize speech.The library "pygame" is used load the module pygame.mixer that contains classes for loading Sound objects and controlling playback. Basically this provide the feature that user give any command in audio form as like "detect","read",etc and our device first convert it in text and them perform the respective task.

8.2 Detailed description

8.2.1 Voice Assistant

The library is used for voice assistant is pyttsx3. Pyttsx is a cross-platform speech (Mac OSX, Windows, and Linux) library. We can set voice metadata such as age, gender, id, language and name. Thee speech engine comes with a large amount of voices for our device we are using dafault voice. When our device call pyttsx3.init() factory function to get a reference to a pyttsx3.Engine instance then it initializes a pyttsx3.driver.DriverProxy object responsible for loading a speech engine driver, then speech engine started and give the desired output in audio format.

8.2.2 Speech Recognition

The first component of speech recognition is, of course, speech. Speech must be converted from physical sound to an electrical signal with a microphone, and then to digital data with an analog-to-digital converter. Once digitized, several models can be used to transcribe the audio to text. In our device two library were used `speech_recognition` and pygame. `speech_recognition` supports many speech recognition engines and APIs, including Google Speech Engine, Google Cloud Speech API, Microsoft Bing Voice Recognition and IBM Speech to Text. And we are using google speech recognition API beacause this API converts spoken text into written text, briefly Speech to Text. You can simply speak in a microphone and Google API will translate this into written text.The API has excellent results for English language. pygame library contain pygame.mixer module that contains classes for loading Sound objects and controlling playback. All sound playback is mixed in background threads. When you begin to play a Sound object, it will return immediately while the sound continues to play. A single Sound object can also be actively played back multiple times.

Chapter 9

Device Interface

9.1 System overview

We have integrated all the 7 modules that is real time object detection, depth estimation, real time face recognition and updation, scene description and image captioning, optical character recognition, emergency video calling service, and speech recognition and voice assistant on our final program. We have integrated all the module in parallel so that one module doesn't effect on speed of other module and for that we use python library "threading" from that we import Thread module that allow us to run in parallel.We have used voice assistant in very efficient way such that device give only required data in audio form. We use single button when user click on that, sppech recognition is on and user give any desired command.

9.2 Detailed description

9.2.1 Speech recognition mode

User needs to press 's' to on speech recognition mode than user can gives any command to our device by his voice. Every time when user wants any mode then he/she needs to first on speech recognition mode.

9.2.2 Detection mode

For detection mode users first need to press 's' then speak "detect" and then the detection mode will be on. In detection mode device will give information in audio form about surrounding objects and also tells person's names by recognizing their faces.

9.2.3 Reading mode

For reading mode users first need to press 's' then speak "read" and then the read mode will be on. In reading mode we will using ocr to extract text from image if present then read mode convert it in audio format and speaks. Basically this mode will come to use when user wants to read any books or articles.

9.2.4 Obstacles mode

For obstacles mode users first need to press 's' then speak "on" and then the obstacles mode will be on. In obstacles mode if any objects or person comes closer approx 0.5 meter then device will speak obstacle has been detected. Basically this mode will come to use when user wants to go any places such that our device prevent them to stuck with any objects.

9.2.5 Caption mode

For caption mode users first need to press 's' then speak "caption" and then the caption mode will be on. In caption mode user will get scene description in audio format. Basically this mode will come to use when user wants know about any place in which he/she present.

9.2.6 Emergency video calling mode

For emergency video calling mode users first need to press 's' then speak "call" and then the emergency video calling mode will be on. In emergency video calling mode user will call to someone else such that he/she can guide user. Basically this mode will come to use

when user stuck any critical situation when our device not able give enough information and user wants help of someone else then by calling to him/her, he/she can see the real time video of kinect and by seeing surrounding condition he/she can guide the user.

9.2.7 Train mode

For train mode users first need to press 's' then speak "train" and then the train mode will be on. In train mode user can add unknown people in dataset and train him/her, so that for next time device can easily recognize his/her face. Basically in this mode we will first create embeddings of unknown person and add this embedding in our file that contains embeddings of dataset, so that for next time device can easily recognize that unknown person.

Chapter 10

Future Improvements

10.1 Scope of Improvement

Major future improvement include using Stereovision over IR rays for depth estimation as it gives better and more accurate results and is faster than previous method. The device will be extended with the functionality of localization for getting the location of person and path tracing for reaching from one location to another.this will be done with the help of GPS+IMU sensor along with Stereovision camera for SLAM technique. Improvements will be done on existing features such as increasing the number of classes of object detection, including audio and other fuctionality to our video calling service. For use at different geographical regions, OCR in that language need to be constructed especially for Hindi. The Prototype needs to be deployed over an embedded system for scalability and usability. Microsoft announcement of new Kinect Azure Camera Sensor is a huge boost for this application as it is more powerful, backed with Microsoft Azure and is very slim and light weight.

References

1. OpenCV Documentation

2. Kinect Documentation

2. pyimagesearch.com

3. stackoverflow.com

4. Falanga, Davide; Zanchettin, Alessio; Simovic, Alessandro; Delmerico, Jeffrey; Scaramuzza, Davide (2017). Vision-based Autonomous Quadrotor Landing on a Moving Platform. In: IEEE/RSJ International Symposium on Safety, Security and Rescue Robotics, Shanghai, 11 October 2017 - 13 October 2017, 1-8.

5. https://blog.miguelgrinberg.com/post/video-streaming-with-flask

6. MAVI: An Embedded Device to Assist Mobility of Visually Impaired

7. www.DeepAI.org

8. Learning CNN-LSTM Architectures for Image Caption Generation by Moses Shos ,Department of Computer Science , Stanford University

9. https://pypi.org/project/pytesseract/

10. https://www.digitaltrends.com/mobile/blind-technologies/

11. https://nei.nih.gov/news/briefs/five-innovations-harness-new-technologies-people-visual-impairment-blindness 12. https://www.goodnet.org/articles/4-innovative-technologies-to-help-blind-people-see-again 13- www.coursera.com-Andrew NG lecture on CNN 14- www.medium.com

15. https://github.com/tensorflow/models/tree/master/research/object-detection 16. https://ieeexplo