Dissertation on

“Object Recognition and Video Analyser for the visually impaired”

Submitted in partial fulfilment of the requirements for the award of degree of

Bachelor of Technology
in
Computer Science & Engineering

UE19CS390B – Capstone Project Phase - 2

Submitted by:

A Spoorthi Alva	**PES2UG19CS001**
K Bharath	**PES2UG19CS189**
Noorain Raza	**PES2UG19CS269**
R Nayana	**PES2UG19CS306**

Under the guidance of

Prof. Swati Pratap Jagdale
Assistant Professor
PES University

June - Nov 2022

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

FACULTY OF ENGINEERING

CERTIFICATE

This is to certify that the dissertation entitled

'Object Recognition and Video Analyser for the visually impaired'

is a bonafide work carried out by

A Spoorthi Alva	**PES2UG19CS001**
K Bharath	**PES2UG19CS189**
Noorain Raza	**PES2UG19CS269**
R Nayana	**PES2UG19CS306**

In partial fulfilment for the completion of seventh semester Capstone Project Phase - 2 (UE19CS390B) in the Program of Study -Bachelor of Technology in Computer Science and Engineering under rules and regulations of PES University, Bengaluru during the period June 2022 – Nov. 2022. It is certified that all corrections / suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 7th semester academic requirements in respect of project work.

Signature	Signature	Signature
Prof. Swati Pratap Jagdale	Dr. Sandesh B J	Dr. B K Keshavan
Assistant Professor	Chairperson	Dean of Faculty

External Viva

Name of the Examiners **Signature with Date**

1. ____________________________ ____________________________

2. ____________________________ ____________________________

DECLARATION

We hereby declare that the Capstone Project Phase - 2 entitled **“Object recognition and Video Analyzer for the visually impaired”** has been carried out by us under the guidance of Prof. Swati Pratap Jagdale, Assistant professor and submitted in partial fulfilment of the course requirements for the award of degree of **Bachelor of Technology** in **Computer Science and Engineering** of **PES University, Bengaluru** during the academic semester June – Nov. 2022. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.

PES2UG19CS001	**A Spoorthi Alva**
PES2UG19CS189	**Koduru Bharath Subba Reddy**
PES2UG19CS269	**Noorain Raza**
PES2UG19CS306	**R Nayana**

ACKNOWLEDGEMENT

I would like to express my gratitude to Prof. Swati Pratap Jagdale, Department of Computer Science and Engineering, PES University, for her continuous guidance, assistance, and encouragement throughout the development of this UE19CS390B -Capstone Project Phase – 2.

I am grateful to the Capstone Project Coordinator, Dr. Sarasvathi V, Professor and Dr. Sudeepa Roy Dey, Associate Professor, for organizing, managing, and helping with the entire process.

I take this opportunity to thank Dr. Sandesh B J, Chairperson, Department of Computer Science and Engineering, PES University, for all the knowledge and support I have received from the department. I would like to thank Dr. B.K. Keshavan, Dean of Faculty, PES University for his help.

I am deeply grateful to Dr. M. R. Doreswamy, Chancellor, PES University, Prof. Jawahar Doreswamy, Pro Chancellor – PES University, Dr. Suryaprasad J, Vice-Chancellor, PES University for providing to me various opportunities and enlightenment every step of the way. Finally, this project could not have been completed without the continual support and encouragement I have received from my family and friends.

ABSTRACT

We propose a robust model that can aid visually impaired people in getting a feel of things around them. Being visually challenged at any level can make it difficult to not only see but also feel what is going on around you. Moving around and doing things on one's own is difficult.

Visual impairment does not always imply total blindness. According to studies, there were over 285 million visually impaired people worldwide in 2012.Most ofthe objects that exist in and around us are of the same shape and size, making it evenmore difficult for a visually challenged person to understand his surroundings. Clinical depression affects nearly one-third of the visually impaired, which is more than twice the general population. This unmistakably drew our attention to it.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Overview

With the massive evolution of technology and the smartphones, we wish to build an effective tool that can assist the blind or visually impaired to overcome their shortcomings and understand their environment better.

Even a slight visionary inability can affect a person’s day to day life hugely. The inability to see the surroundings makes it difficult to move around or do anything on one’s own.

Visually impaired are our major user class, being our primary communication actor with this interface, would be prompting the application with completely voice enabled features. Any person with even the slightest disability in viewing the surroundings can use this application.

1.2 Purpose

It is estimated that there are approximately 36 million visually impaired people in the world. The World Health Organization indicates that up to 80% of global visual impairment is preventable through improved access to treatment, but as the world's population ages, the number of people who are blind or partially sighted will increase. In light of this information, our aim is to propose a robust model which can aid visually impaired people to get a feel of things around them in the form of an android application. Our application is built to meet a user being a visually impaired person, who is eager to know his surroundings without anybody’s help. Our proposed system aims to have an object recognition model along with video analysis based on the object chosen by the user. Modularizing this approach enables video to speech conversion as well.

1.3 Scope

Our project covers visually impaired or blind people as a scope of the project. Our proposed system aims to have a distance estimation model along with video analysis. Modularizing this approach enables video to speech conversion as well.

CHAPTER 2

PROBLEM STATEMENT

Our proposed system aims to have an object recognition model along with video analysis based on the object chosen by the user. Modularizing this approach enables video to speech conversion as well. Benefits: Whenever a visually impaired person wants to understand his/her environment better he must look for help, we, in our model try to eliminate this by bringing in a robust model which can understand the surrounding better and give the user a description from the perspective of the object needed. Objectives to achieve: completely voice automated system. automated recording setup multiple object recognition model.

- Real time distance estimation
- Video description to be provided
- Description in the form of audio

CHAPTER 3

LITERATURE REVIEW

3.1 Guiding visually challenged over a smartphone using CNN

3.1.1 Introduction

This app directs the visually impaired to the object using voice feedback. Major novelty here is using the user’s hand as a reference object for metrics analysis.

3.1.2 Characteristics and Implementation

The system detects the user's hand as it enters the range of perception of the camera and recognises the other target objects by using CNN for object detection, a single shot multibox detector approach position is estimated as directions using machine learning and image processing techniques. Tensor flow lite, was trained on the COCO dataset.

3.1.4 Features

In this paper, CNN for obj detection and SSD approach are used. SSD outperforms YOLO and all other CNNs in terms of frequency per second accuracy. The COCO dataset was used to train the SSD-based TensorFlow Lite model. This model's workflow is as follows: the object to be found is given as input into the system, the model recognises the hand to calculate the distance of reach, and then gives outputs.

3.1.5 Evaluation

Limitation: Wong et al. have already developed a model based on the Caffe framework that uses a CNN model for object detection.

This model also connects to the Azure cloud platform, allowing for remote access. This is unquestionably a breakthrough in the field.

Benefits: The object is not required to be in a specific location. The most common model necessitates that the object be in a specific position. A nonlinear model capable of learning low and high level skin features is included.

3.2 Guiding for visually challenged using windowing based mean method on Microsoft Kinect camera

3.2.1 Introduction

This is an obstacle avoidance system based on the Xbox 360's Kinect depth camera. It employs a window-based approach that emphasizes voice feedback. To read the coded light black-SoC chip, the infrared rad-cam, for example, employs a CMOS receiver.

3.2.2 Characteristics and Implementation

Images are preprocessed to increase processing speed by converting them from 16 bit unsigned int to 8 bit unsigned int. Any salt and pepper noise caused by errors is removed by the median filter. Object detection using a novel type windowing technique. The brightness of an object as it is approached is used to calculate distance.

3.2.3 Features

Benjamin invented a laser cane that produces beams of infrared light. The reflected light is received by the photodiode via the receiver lens.. To estimate how far the object is from user ,the triangulation method is used.

3.2.4 Evaluation

Advantages: This proposed system has a wider range (4m).Disadvantages, in this regard, Wahb introduced a system that includes an ultrasonic sensor, microcontroller, vibrator and water sensor**.**

3.3 Smart personal Artificial intelligence aid for visually impaired

3.3.1 Introduction

Intelligent retrieval and machine intelligence are the most rapidly evolving technologies. These automations are critical to the advancement of the information technology sector. This paper attempted to use these technologies to help visually impaired people live independent and normal lives.

3.3.2 Characteristics and Implementation

The REST API given by the cloud application interface, which encapsulates machine intelligence, is used to analyse the captured image. It quickly categorises images (for example, "Taj Mahal," "Deer," and "Footwear"), detects individual features in faces within images, points and reads textual words within images

3.3.3 Features

This white paper discusses REST APIs and how they can benefit the visually impaired. The workflow includes voice inputs, NLP, textual data to the DialogFlow server, textual analysis, and voice response. The image capturing workflow is as follows: capture image from camera, upload image, image analysis with vision API, and server response.

3.3.4 Evaluation

The Vision Application interface displays the output screen after analysing the pictures captured by the camera. As a result, the image is analysed based on the image category, and JSON data with various image parameters is returned to the app. The enjoyable conversation with the bot identification of objects and their environment in images to facilitate payment, currency recognition is used.

3.4 Design and Implementation of Voice Assisted Smart Glasses for visually Impaired People Using Google Vision API.

3.4.1 Introduction

In general, visually impaired people find it difficult to manage many types of challenges in everyday life, including traveling. Most wooden sticks are used to find obstacles and barriers next to them.

Because of this, the visually impaired cannot know exactly what challenges they face and must rely entirely on the lead her stick and training to navigate safely in the right direction. Research in this area focuses on developing guidance systems that use a combination of smart glasses and sensors to continuously acquire images from the environment through wearable smart glasses.

3.4.2 Characteristics and Implementation

In a real-world scenario, the authors of this research used an approach to discover the direction among many straight paths in varied surroundings. When the image is acquired, these straight pathways will emerge for converging to a predefined position from the vanishing point and will have the matching description. The examination of these parallel edges is aided by the suitable isolation and practical simulation of the trapped frame. After that, the vanishing point is calculated, and a decision-making system is built to indicate the visually impaired individual his or her deviation from the straight line. The Darknet is a highly complicated Convolutional Neural Network architecture using the You Just Look Once (YOLO) algorithm. YOLO uses a convolutional neural network (CNN) for real-time object recognition.

3.4.3 Features

The smart glass has a processor that analyses and recognises objects in order to inform the viewer of the image's results so that they have a comprehensive picture.It's a speech-based user interface in which the user speaks a voice that interprets his destination location as he approaches.This strategy can be used to locate an obstruction in a specific location. After the obstacle has been detected, the user will be sent to their destination. As a result, the case for having two borders on the ROI's left and right margins is strong.

As a result, the return on investment will be divided into three pieces. The user will be able to pass if the obstacle is on the right side.

3.4.4 Evaluation

The software can recognise and take into account a variety of risks that may be encountered when driving, such as common objects and machinery, various types of vehicles, food, and so on. An impediment can be located in a specified location using this method. The user will be guided to their destination after the barrier has been identified.

Experiments are carried out over a variety of distances ranging from one to fifteen metres. In addition, as an optional feature, the user can instruct the system to cease vibrating by speaking a voice command such as "Stop Vibration" until contact is made with the object. A variation in acceleration also occurs when the destination is the farthest away from the nearest point. With decreasing distance between the obstacle and the user, the vibration type changes from short to continuous, indicating to the user a constant alarm about dangerous conditions

3.5 Image Captioning based Smart Navigation System for Visually Impaired

3.5.1 Introduction

Blindness is a major condition that affects people all around the world. It not only has an influence on the patient's life, but it is also one of the major causes of the global financial crisis, according to the Lancet Global Health report[2020]. Apart from that, it has the potential to limit an individual's freedom as well as their ability to complete tasks to which they are capable, resulting in unemployment and financial insecurity. Visually impaired people seem and act much like the rest of us, except they can't

see. Individuals who have poor vision or who have good vision seek assistance. For a visually impaired person, self-navigation in novel outdoor conditions or travelling around a known site are daily challenges.

3.5.2 Characteristics and Implementation

A tiny spy camera would be placed in the centre of the spectacles, capturing real-time photographs. A Type-B charging port will be included in the specifications. The activation buttons for power and read mode are also provided. To show whether the buttons are switched on or off, basic indicators are given.The photographs will be uploaded to the cloud and fed to the model, who will provide a verbal output. The spectacles come with a Bluetooth module that connects to the headphones and provides voice output to the user.

3.5.3 Features

In this study, we propose a novel implementation of smart spectacles based on Image captioning and Optical Character Recognition (OCR).There is an embedded camera, an OCR module, a text-to- speech module, and an Image Captioning module. Furthermore, it allows blind people to comprehend signboards and utilize the available facilities to the fullest extent.

Using the presented model, captions are successfully generated based on the surrounding environment and the text can be read whenever required.

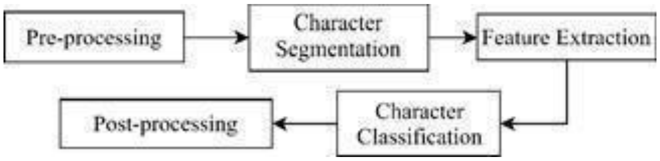


Fig1: Working of OCR

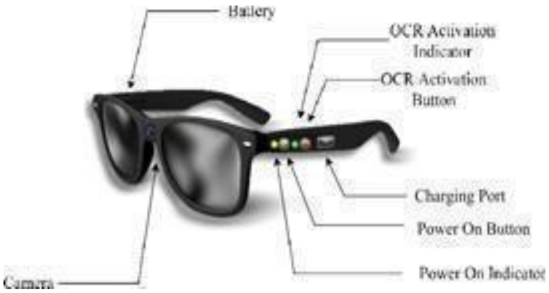


Fig 2: Design of spectacles:

The image captioning model consists of:

- Encoder-decoder model
- Word embedding model
- Beam search

3.5.4 Evaluation

In real time, the model converts the image to a textual representation. Text is represented as a picture containing text or an image depicting a scene. This textual output is subsequently converted to speech in order to provide notifications to the user. The future of this proposed model can be linked to Google Maps, allowing text to be broadcast to other devices in the network for analysis. As a result, it will become more powerful and intelligent. This research shows how various technologies can be used to produce a powerful system that can aid in the management of substantial life impairments.

3.6 Speaker disentanglement in video to speech conversion

3.6.1 Introduction

The purpose of this research is to learn a mapping from a silent lip video of a person talking to its corresponding audio speech signal while also managing the speaker identification of the output speech. This job has a compelling application in that it allows people who have lost their ability to speak to communicate with a speech synthesizer in a more personalized, fast, and natural manner. This method has the advantages of being able to manage the speaker identification at inference time and (ii) being able to use larger quantities of training data by permitting several speakers to be included in the dataset.

3.6.2 Characteristics and Implementation

The purpose of video-to-speech is to convert a silent video of lip movement into an audible signal. Previously, most approaches to this problem were limited to a single speaker, but this method allows for numerous speakers. To improve input conditioning and make the model less dependent on the autoregressive signal, the authors of the research used a dropout-like method. The idea is to replace a fixed value with a variable value.

3.6.3 Features

This research focused on the task of multi-speaker video-to-speech conversion. It enables persons who have lost their ability to communicate to communicate through a voice synthesizer in a more personalized, fast, and natural manner.

The audio decoder then receives the improved vector, which combines the speaker identification with the visual information (obtained after the LSTM).

It is divided into two parts:

- The use of video-to-speech synthesis is being investigated.
- It aids in the regulation of the speaker identity of the produced audio. The GRID corpus was used in the experiment. The fact that it is not a real-time programme is one of its drawbacks.

3.6.4 Evaluation

For the job converting video to speech, we compare our proposed methods to previously proposed alternatives. We have a look at the speaker-dependent configuration, which consists of four speakers with 900 training samples, 50 validation samples, and 50 testing samples apiece. We put our methods to the test in three different ways: a speaker- independent baseline trained on all four speakers at the same time, a speaker-dependent baseline learned on each speaker separately, and a model trained on all four speakers at the same time but explicitly integrates the speaker identity (SI). The numeric data is shown in Table I, while qualitative samples can be viewed online.

3.7 Eye Assistant : Using mobile application to help the visually impaired

3.7.1 Introduction

It is a product that is based on an app that recognises objects and text. The Machine Learning framework is used to detect any object, followed by the Google Text Speech API.

3.7.2 Characteristics and Implementation

In this dataset, CIFAR 10-600 images and 10 objects were used. Using a machine learning framework and the Google Text to Speech API, it detects any object. It is not necessary to use a remote server.

Voice feedback is provided as an output. It scans the environment rather than taking photographs for processing. Tensor flow is used to eliminate background noise. It also works with low-end smartphones

3.7.3 Features

It does not necessitate the use of a remote server. It is not necessary to save the photographs. OCR is used to recognise text.

3.7.4 Evaluation

Disadvantages include the fact that it is incompatible with low-end smartphones. Advantages: Other available applications are expensive.The majority of existing applications are concerned with currency detection.

3.8 Real-Time Text Tracking for Text-to-Speech Translation Camera for the Blind

3.8.1 Introduction

This study proposes a reading helper device prototype with a locator that focuses on the textual elements present in the scene which uses sound signals to show the text location which also includes a text tracking which occurs in real- time to extract textual information.

This divide basically acts as a personal assistant to read out the textual information and signs that is present around the visually impaired person to help him navigate the surroundings in a better way.

3.8.2 Characteristics and Implementation

The scene text finder, OCR model to capture all the details in the imputed image and then a text - speech model to convert the textual information captured to audio format so that the user can hear it. All these medals were combined to create a prototype device.

3.8.3 Features

It utilizes a wearable camera and its a real time model for recognizing textual features in a given frame captured by the camera. and the output is in the audio format which the user can hear through a earphone/headphone.

3.8.4 Evaluation

This can handle occlusion which occurs temporarily in the scene that can occur as we cant predict the surrounding all the time.It can also handle text regions that is either slighlty out of view or blurred. Only provides conversion of the textual regions of images captured into audio conversion.

3.9 Providing Synthesized Audio Description for Online Video

3.9.1 Introduction

This research paper aims to provide a standard software for including audio caption to online videos found on the internet so a visually impaired person can enjoy those videos to an extent.

3.9.2 Characteristics and Implementation

An editor to edit a script, a video platform which plays the video, a repository which contains all the metadata and a protocol which is completely text based is used for exchanging audio captions or description scripts between the parts to make up solutions provided in this research paper. it utilizes a technique to use synthesized speech to provide the audio caption.

3.9.3 Features

It uses synthesized speech to include audio captions. Works on any online video playing and sharing services like news, youtube etc.

3.9.4 Evaluation

It is limited to translating prerecorded online videos and human intervention is needed to translate it and no algorithms model has been us.

3.10 Automated Video Description for Blind and Low Vision Users

3.10.1 Introduction

This is basically a description with an additional query model for visually impaired people to properly understand what's happening in a video being played and this model makes it more interactive as a person can pose questions related to objects on demand. The aim is to create an interactive model which keeps human interactions other than a blind person to a minimum.

3.10.2 Characteristics and Implementation

The paper talks about two prominent models that are NarationBot and Info bot. NarrationBot is implemented using several models such as YOLOv3, CNNs and RNNs, Pythia caption generation model, ImageNet - (OCR) api and Text Summarization to generate general description for the scene that is occurring in the played video. InfoBot is a visual dialog model which uses m- regional convolutional neural network for each picture which accumulates the extracted attributes or features and provides it to LSTM and text- audio API to answer all natural language questions possessed by the user. It uses COCO dataset to train the models. The events to trigger the events here are keyboard keys. When the video is being played the user can pause the video by using the key ‘Q’ and then ask his/her queries and back an answer in real time.

3.10.3 Features

It eliminates external human interaction as a visually impaired person alone can facilitate its functionalities without taking external sources help. They have uses different conditions to evaluate which combination of models works best.

3.10.4 Evaluation

Information is not always accurate and has a small sample size associated with it. It's specifically designed for online videos to be played on a system with keyboards and uses keyboard keys for initialing the audio output. But it provides on-demand description which we wish to incorporate in our model.

CHAPTER 4

PROJECT REQUIREMENT SPECIFICATION

4.1 Operating Environment

The operating system used would be any android version, preferably android 11. Hardware platformis smartphone. Software component: The application with the model as back-end.

4.2 Regulatory policies

As we are developing an app, the application should be such that it does not interfere with the working of the other applications

4.3 Hardware limitations

The user must carry a charged phone. Cloud storage should be sufficient to store the videos after recording if given the option. A functioning microphone, and speaker should be essential. The input commands must be clear. The camera that the phone uses should be of sufficient megapixels so that the object recognition model can recognize the attributes clearly.

4.4 Limitations of simulation programs

Deployment of a functioning app may be difficult. The delay between the input and output generated would impact on the experience of the user. Network instability might be a hindrance. As the audio output is in English, the maximum utilization of the app is only beneficial to a user who is fluent in English language.

4.5 Interfaces to other applications

As our sole focus is in developing an app, our model does not work on a web based platform.

4.6 Parallel operations

The audio output of the object analysis takes place as in when the video analysis is processing the input.

4.7 Criticality of application

Processing can be limited to certain valid inputs. There might be sensitive information disclosure. Information leakage of the user stored in the phone and Insufficient input validation might be an issue.

4.8 Risks

A stable internet connection is essential. Slight delay in the processing of the input. Low-end hardware which compromises the app to fully deliver to its potential.

4.9 Functional Requirements

This model would be in the form of an application which could be compatible with any low-end systems as well. We aim to look from the perspective of the user to have a more relevant and efficient interface,in terms of the functionalities to be offered. The application would run with the help of voice command.As soon as the app opens there will be a recording window for taking the video as input. After the said window, the model would start processing the video.

The model should be able to analyze various objects from the video input. It should prompt the user to input an object. The model would generate a description based on the input provided. It should analyze the video with respect to the object given as input by the user. In case the user does not give any specific object details to be retrieved then, the model should be able to give a generalized description of the video. All the input given to the user would be in the form of audio. So, the model requires video to audio conversion. Since the distance estimation feature also exists, we have to do careful consideration on all aspects of measurements.

4.10 External Interface Requirements

4.10.1 User Interfaces

Tkinter is the standard GUI library for Python. Python when combined with it provides a fast and effortless way to create GUI applications. Tkinter provides a powerful object-oriented interface to the Tk GUI toolkit. Screen layout would include a basic User interface.

The application would completely voice integrated. The input being a video is completely timed. After the said seconds, the processing would be activated.

4.10.2 Hardware Requirements

The software works on Android devices. It uses the microphone and the speaker to perceive and give the output.

4.10.3 Software Requirements

Datasets used:

- COCO (Common Objects in Context) dataset
- Flickr 32k dataset

You only look once-It applies a neural network to the entire image to predict bounding boxes and their probabilities. High probabilities images are considered as objects detected. Pretrained YOLOv3 can detect 80 objects such as a person, car, bike etc.

Transfer learning: A technique to reuse weights in one or more layers from a pre-trained network model for a new model by preserving the weights, fine-tuning the weights, and fully adjusting the weights when training the new model.

Multiple object detection using tensor flow: Object Detection using TensorFlow is a computer vision technique. As the name suggests, it helps us in detecting, locating, and tracing an object from an image or a video.

OpenCV- Object recognition is a computer vision technique that enables software systems to detect, locate, and track objects based on specific images or videos. A special aspect of object recognition is the identification of classes of objects (people, tables, chairs, etc.) and their site-specific coordinates within a given image. Position is indicated by drawing a bounding box around the object.

The bounding box may or may not precisely locate the object. The ability to find objects in images defines the performance of algorithms used for detection.

The Python Imaging Library (PIL) adds image processing capabilities to your Python interpreter. This library supports many file formats and provides powerful image processing and graphics capabilities.

Python text to speech library is available in python which helps is text to speech conversion. Flask is a micro web framework written in Python and based on the Werkzeuge toolkit and Jinja2 template engine.

Python tensor flow for distance calculation makes use of standard datasets, which help train our model and make it accustomed to adapting using the latest AI tools to get the desired results.

4.10.4 Communication Interfaces

The communication Interface is the internet available for the user such as the Wi-fi, LTE, hotspot etc. The line speed can be over 2mbps to have a stable and reliable connection.

4.11 Non-Functional Requirements

4.11.1 Performance Requirement

The product perceives the input of the surroundings and subsequently implements the algorithms to analyze the objects and identify the necessary items. It precisely identifies the objects in the frame anddescribes them to the user when asked for. The person can be completely reliant on the product in any case of any emergency. They can share his location and other useful information with any priority contacts and be safe all they areout travelling.

4.11.2 Safety Requirements

Safety measures that must be taken are that they always have to have a battery back-up with them so that they never run out of charge.

4.11.3 Security Requirements

The user can also authenticate themselves to use the app, however the user just has to login only once while they installed the app and does not need to do it all the time.

4.12 Other Requirements

Scalability can be done for memory management as the memory can either be saved or discarded once the object identification is done.

There is a possibility that there can be a maintenance break in the future if required to adapt to new surroundings. Since the application makes use of the cloud, scalability might not be an issue.Dataset training also works regarding varied kinds of image

CHAPTER 5

HIGH LEVEL DESIGN

5.1 Current system

The current systems have enhanced models which use Yolo and SDD. It has a faster CNN which optimizes the video analysis and the object detection. For video description, we have used two machine learning models. VGG16 for feature extraction and LSTM(Long short term memory) for captioning. LSTM is trained with the Flickr30k dataset, which consists of about 30,000 thousand images with five captions each. First the required frames are extracted using the OpenCV library and then features of those images are extracted using the VGG16 model. Using those features, the trained model is used to predict the caption and an audio description is played.

The next main feature is to analyses the distance of the object with respect to the camera. We have used an algorithm where YOLO can be used to predict absolute distance of any person using the data taken as input from a mobile camera.This feature serves as a threat detector incase the person is too close to the user.

5.2 Design Considerations

5.2.1 Design Goals

The proposed methodology primarily works with three important models along with a front end for easy operability. We aim to address the major feedback mechanism that a blind person would need in case he wants to know about his surroundings. This application has a basic front end where there is a login window, if the user wants to save his details. There is a scan feature which helps the user to record his surroundings and would give the list of objects detected along with the description of what is happening in the surrounding. Along with this if there is any human detected very close to the user, then he would be warned about the distance too. Hence , we have a model which serves all these features for the visually impaired who find it difficult to understand their surroundings better .

5.2.2 Architecture Choices

We have opted for User centric design approach and Bottom up design approach.

In User centric approach, the focus is on putting users at the center of product design and development. We also aim to satisfy the user's needs and want it to be our priority. Since the audience for our project is primarily visually impaired, who can use a basic smartphone, we aim to help these individuals understand their surroundings better without any external help. User centric approach throws light at research, empathy and iteration.

We observe that this model required us to develop root level models primarily and then assemble all of it into one. We have taken utmost care to consider all the minute details that could help us in building this project. Having bottom-up approach ensures that:

- Individual parts of the system are specified in detail.
- Ensures minimum data redundancy and focus is on re-usability.

The parts are linked to form larger components, which are in turn linked until a complete system is formed.It processes incoming data from the environment and creates simple systems.

5.3 Constraints, Assumptions and Dependencies

5.3.1 Assumptions

We assume that our target audience is the visually impaired section of our society who find it hard to understand their surroundings without any external help.

The user has a smartphone which has basic functionalities to record the environment and also has internet connectivity. The user wants to get a simple one line description of his surroundings and information regarding the distance.

5.3.2 Interoperability requirements

The communication Interface is the internet available for the user such as the Wi-fi, LTE, hotspot etc. The line speed can be over 2mbps to have a stable and reliable connection. The user can also make satellite calls based on the mobile they are using. It is the primary source of communication for the user.

5.3.3 Interface/protocol requirements

Tkinter is the standard GUI library for Python. Python when combined with it provides a fast and effortless way to create GUI applications. Tkinter provides a powerful object-oriented interface to the Tk GUI tool Screen layout would include a basic UI. The application would be completely voice integrated. The input being a video is completely timed. After the said seconds, the processing would be activated

5.3.4 Data repository and distribution requirements

The data repository used MICROSOFT COCO model which comprises 80-90 classes and 2 lakh labeled images. Flickr 32k dataset

5.3.5 End-user environment

The user would be using a mobile interface which would specifically be running on an android operating system. The application would completely be voice enabled along with the features being available to the user through just voice commands.

CHAPTER 6

LOW LEVEL DESIGN

6.1 Design Description

We have used different modules which cater to the needs of the different aspects or sub-parts. The first part is the object detection and we have used YOLO v-5 to identify objects

6.1.1 Object detection

YOLO (You Only Look Once) is an object detection approach. It is the fastest hence it is highly recommended for the project as the processing time can be a major concern.

YOLO uses a grid pattern to identify the objects and classify the same. Non-Max Suppression algorithm is used in the methodology as it helps us not to identify the same objects again and again in an image.

6.1.2 Distance estimation

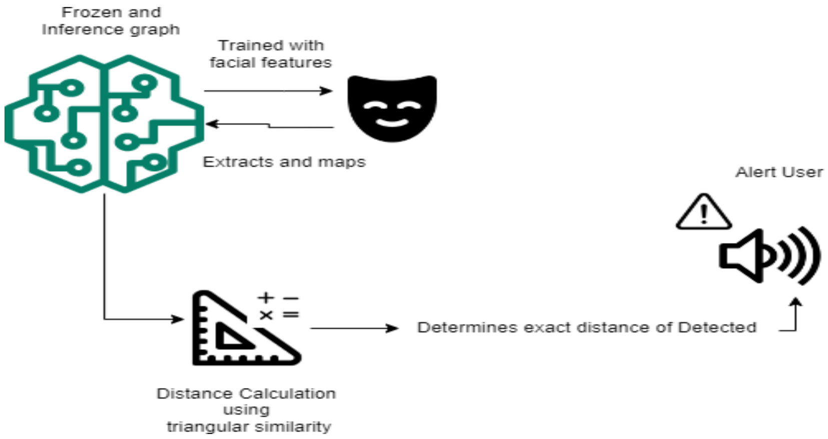


Fig3: Architecture of distance module

We have a distance estimator module that uses Python TensorFlow to calculate the absolute distance of an object relative to the camera. Use triangle similarity to determine the distance from the camera to a known object or marker. Using the triangle similarity, the triangle similarity is:

Suppose we have a marker or object of known width W. Then place that marker at a distance of D from the camera. Take a picture of the object with a camera and measure its apparent width in pixels P. From this we can derive the perceived focal length F of the camera:

$F = (P \times D) / W$

We have functions like find contours to identify the boundaries of an object in a frame. We are making the assumption that the contour with the largest area is our piece of paper.

6.1.3 Video description

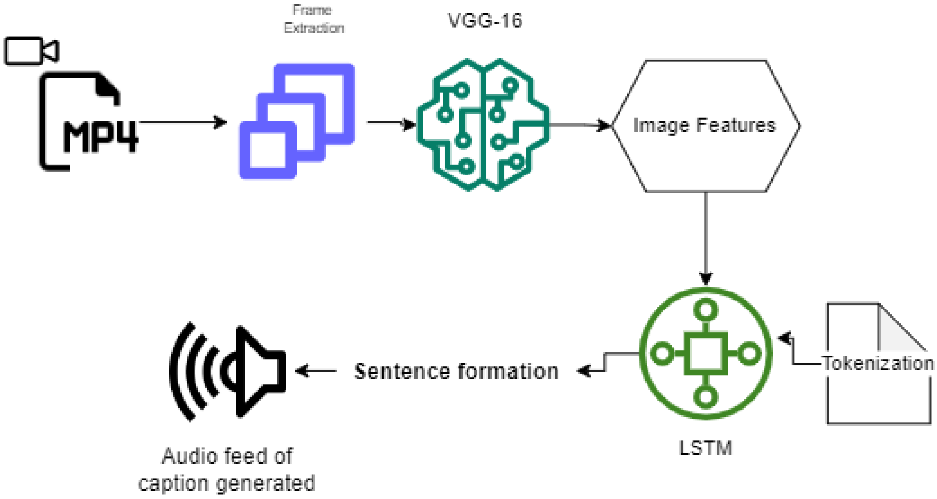


Fig4: Architecture of video caption module

The caption generator uses two machine learning models to extract features and then another for text based processing. The dataset used here is Flickr 30k dataset which has about 31000 images with 5 reference sentences each describing each image. For the image extraction, VGG(Visual Geometry Group)-16 is the first model used for feature extraction of the input image. Then it's saved using the pickle library. Then we load the captions data and map the correct set of captions to the right image by assigning an “id” to each of them. Then preprocessing of the captions takes place and we split the data into training and testing sets. Now the second model that we use is the LSTM(Long short term memory) network for natural language processing.

As the dataset is quite large we have to create batches and then train the model so that the system doesn't crash. The Bleu score obtained was 0.55 with epochs set at 20.

After successfully training the model, it successfully predicts the caption for a new image by converting the predicted index to a word, and the caption generator appends all the words which is then fed into pyttsxs3 for audio conversion of the text generated. For a video input, relevant images are extracted using OpenCV and then passed through VGG16 and the trained model to predict the caption in audio format.

6.2 Class diagram

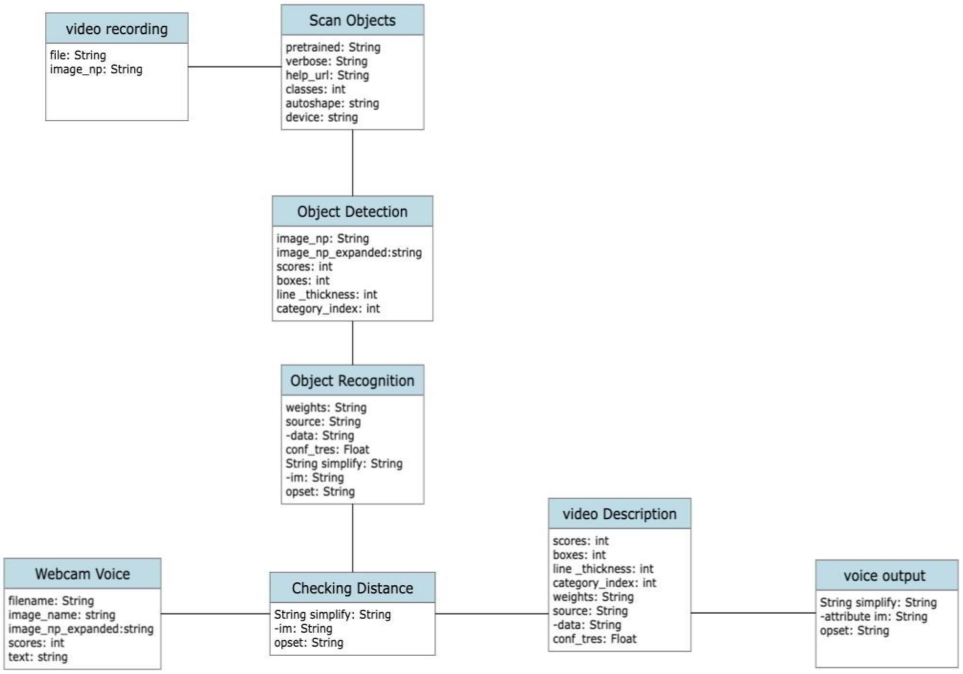
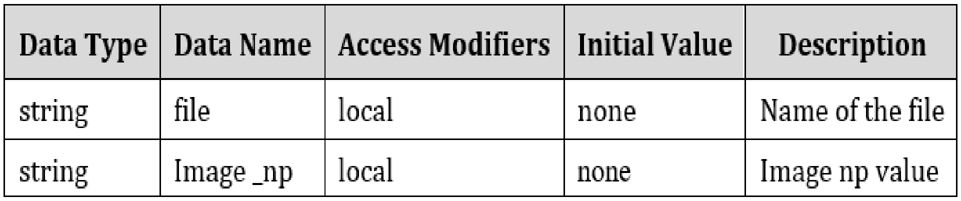


Fig 5 :Class diagram of the model

6.2.1 Class name : Video Description

Class description : To analyse video and predict

Table 1: Data members of Video Description class

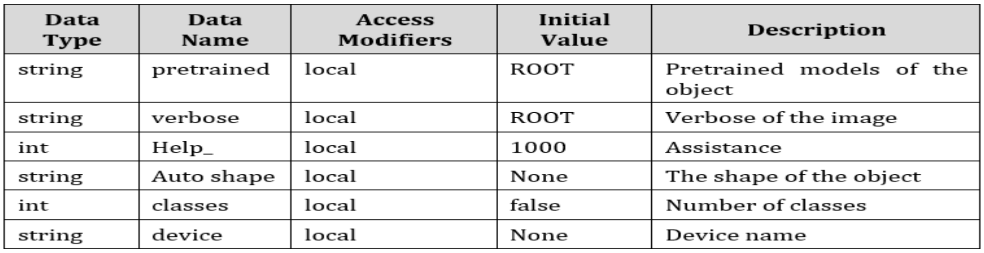


Data Type	Data Name	Access Modifiers	Initial Value	Description
string	file	local	none	Name of the file
string	Image _np	local	none	Image np value

6.2.2 Class name : Scan Objects

Class description : To scan objects

Table 2 :Data members of Scan objects class

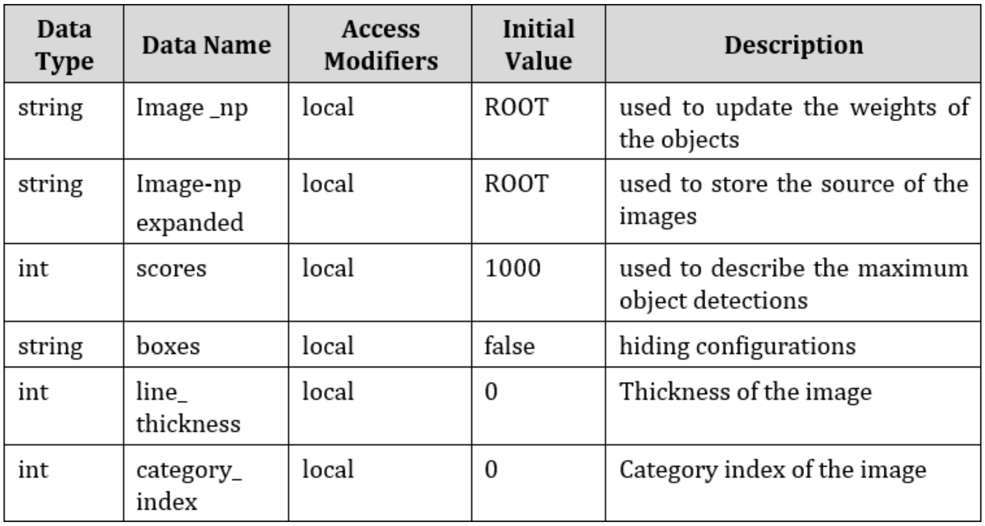


Data Type	Data Name	Access Modifiers	Initial Value	Description
string	pretrained	local	ROOT	Pretrained models of the object
string	verbose	local	ROOT	Verbose of the image
int	Help_	local	1000	Assistance
string	Auto shape	local	None	The shape of the object
int	classes	local	false	Number of classes
string	device	local	None	Device name

6.2.3 Class name : Detecting objects

Class description: Detect objects present in the scanned window.

Table 3 :Data members of Detecting objects class

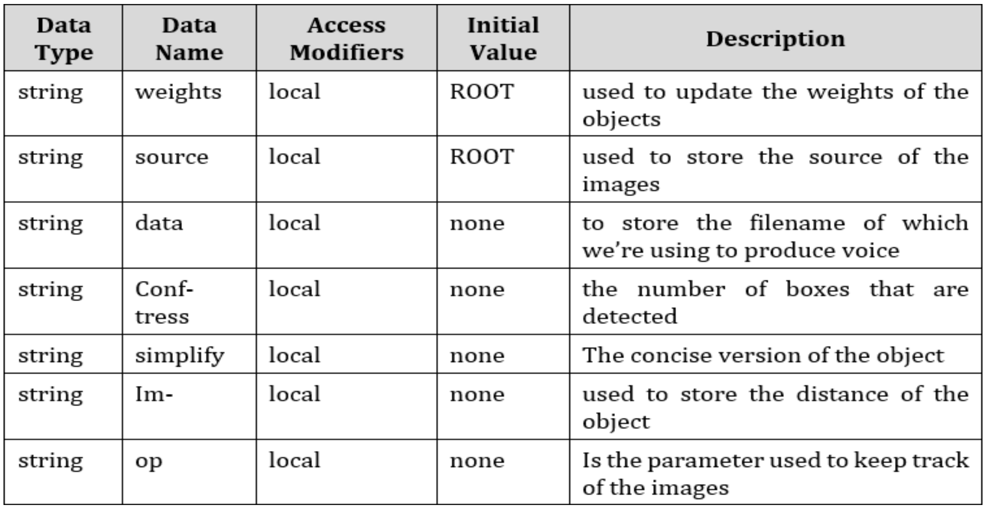


Data Type	Data Name	Access Modifiers	Initial Value	Description
string	Image _np	local	ROOT	used to update the weights of the objects
string	Image-np expanded	local	ROOT	used to store the source of the images
int	scores	local	1000	used to describe the maximum object detections
string	boxes	local	false	hiding configurations
int	line_ thickness	local	0	Thickness of the image
int	category_ index	local	0	Category index of the image

6.2.4 Class name : Object Recognition

Class description: Recognize the objects present.

Table 4 :Data members of Object Recognition class

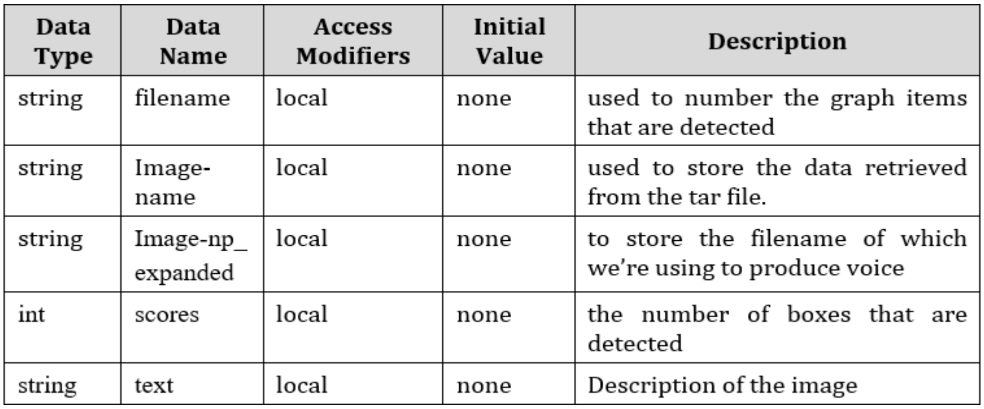


Data Type	Data Name	Access Modifiers	Initial Value	Description
string	weights	local	ROOT	used to update the weights of the objects
string	source	local	ROOT	used to store the source of the images
string	data	local	none	to store the filename of which we're using to produce voice
string	Conf-tress	local	none	the number of boxes that are detected
string	simplify	local	none	The concise version of the object
string	Im-	local	none	used to store the distance of the object
string	op	local	none	Is the parameter used to keep track of the images

6.2.5 Class name : Web cam voice

Class description : Voice integration

Table 5 :Data members of Webcam voice class

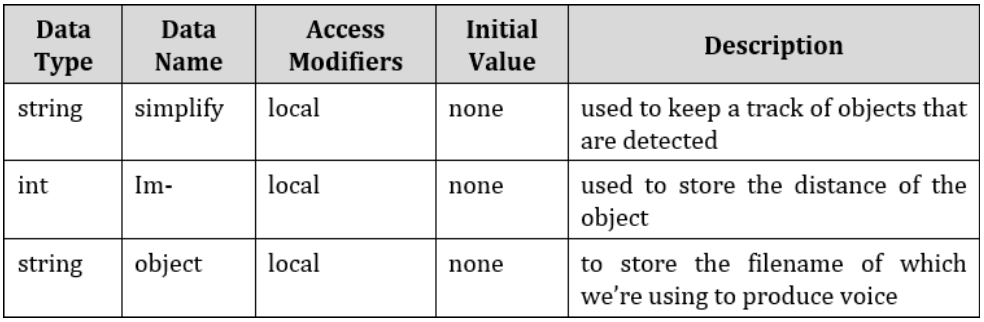


Data Type	Data Name	Access Modifiers	Initial Value	Description
string	filename	local	none	used to number the graph items that are detected
string	Image-name	local	none	used to store the data retrieved from the tar file.
string	Image-np_ expanded	local	none	to store the filename of which we're using to produce voice
int	scores	local	none	the number of boxes that are detected
string	text	local	none	Description of the image

6.2.6 Class name : Checking distance

Class diagram: Tracking the objects

Table 6 :Data members of Checking distance class

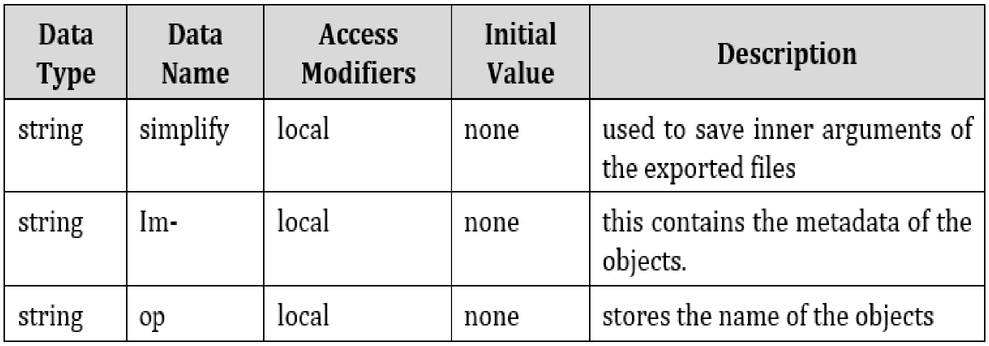


Data Type	Data Name	Access Modifiers	Initial Value	Description
string	simplify	local	none	used to keep a track of objects that are detected
int	Im-	local	none	used to store the distance of the object
string	object	local	none	to store the filename of which we're using to produce voice

6.2.7 Class name : Voice output

Class diagram : Give voice output

Table 7 :Data members of Video Description class

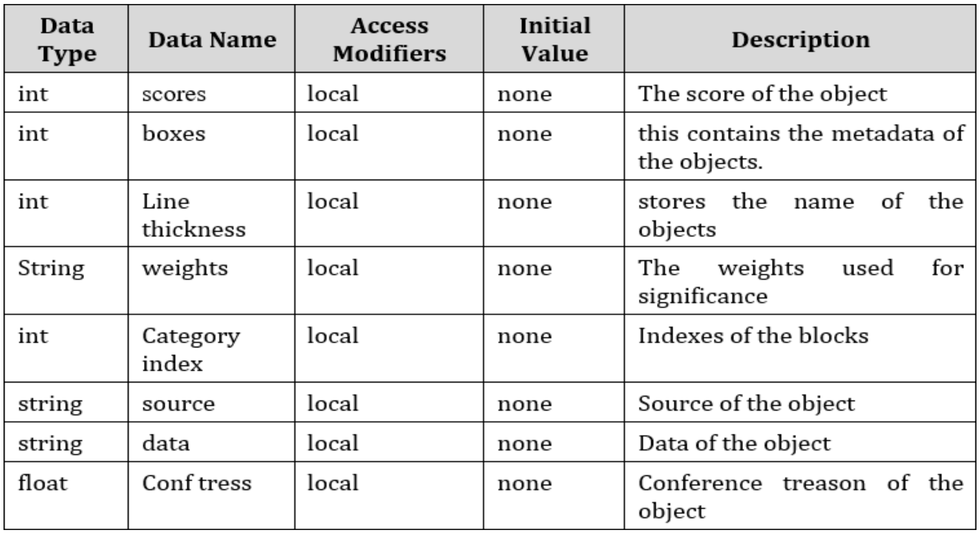


Data Type	Data Name	Access Modifiers	Initial Value	Description
string	simplify	local	none	used to save inner arguments of the exported files
string	Im-	local	none	this contains the metadata of the objects.
string	op	local	none	stores the name of the objects

6.2.8 Class name : Voice description

Class description: Description given in audio format.

Table 8 :Data members of Voice description class



Data Type	Data Name	Access Modifiers	Initial Value	Description
int	scores	local	none	The score of the object
int	boxes	local	none	this contains the metadata of the objects.
int	Line thickness	local	none	stores the name of the objects
String	weights	local	none	The weights used for significance
int	Category index	local	none	Indexes of the blocks
string	source	local	none	Source of the object
string	data	local	none	Data of the object
float	Conf tress	local	none	Conference treason of the object

CHAPTER 7

SYSTEM DESIGN

7.1 USE CASE DIAGRAM

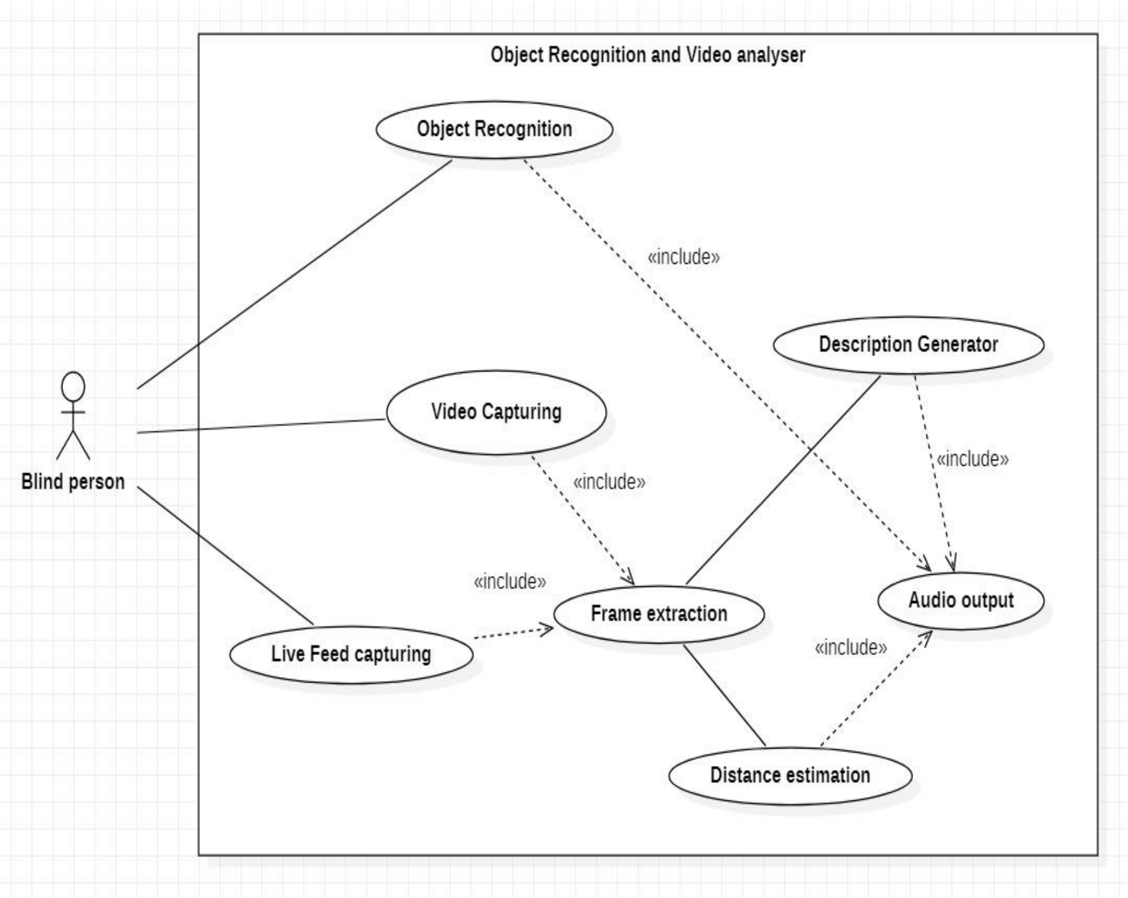


Fig 6 : Use case diagram

7.2 ACTIVITY DIAGRAM

7.2.1 Distance Estimator

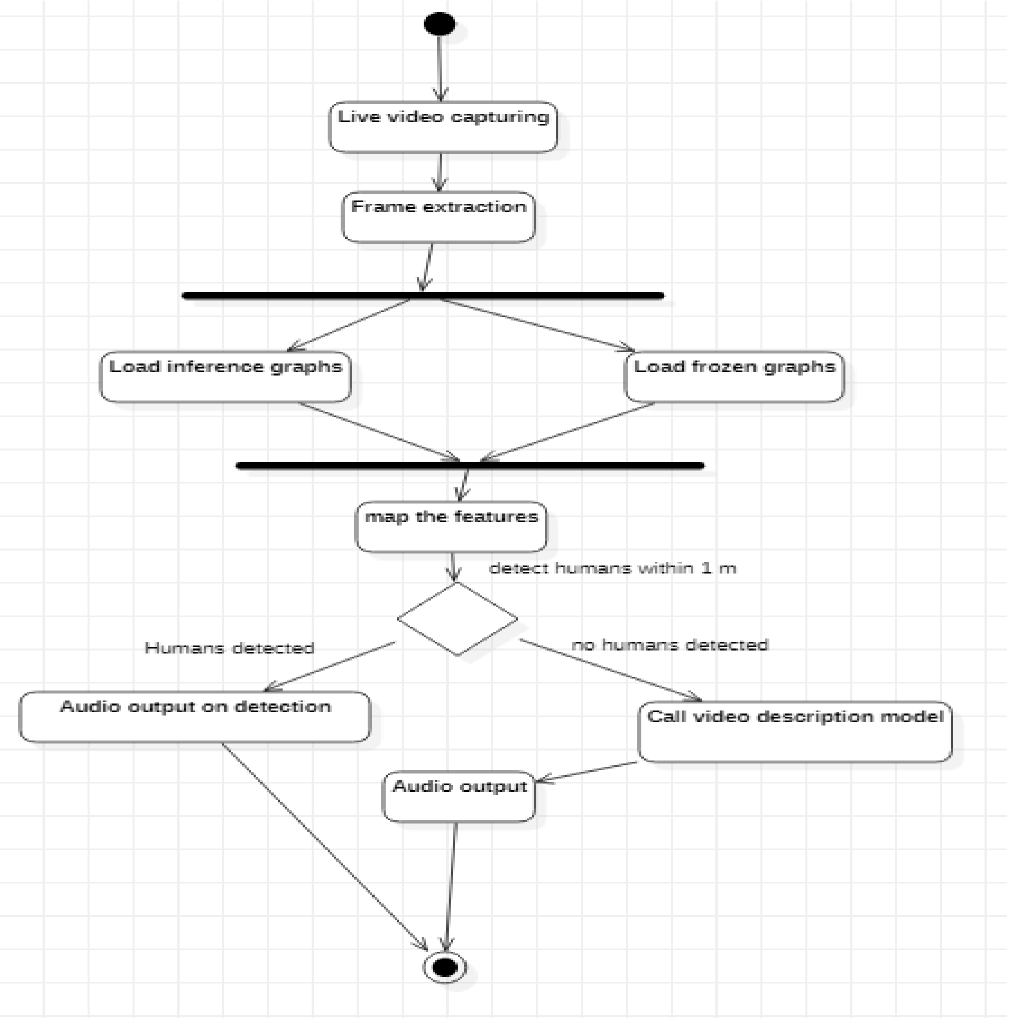


Fig 7: Activity diagram of Distance estimator

7.2.2 Object recognition model

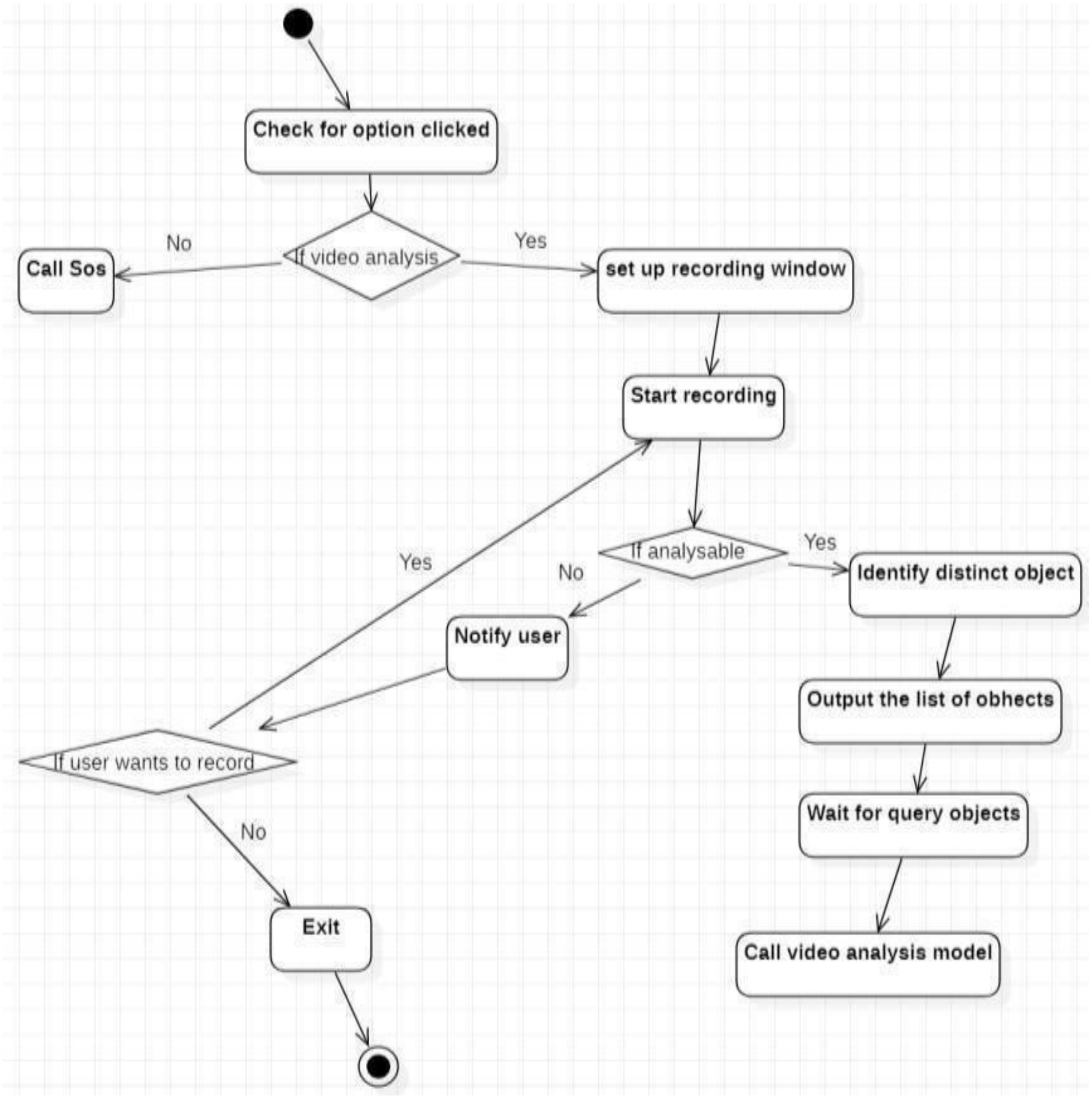


Fig 8:Activity diagram of object recognition model

7.2.3 Video analysis

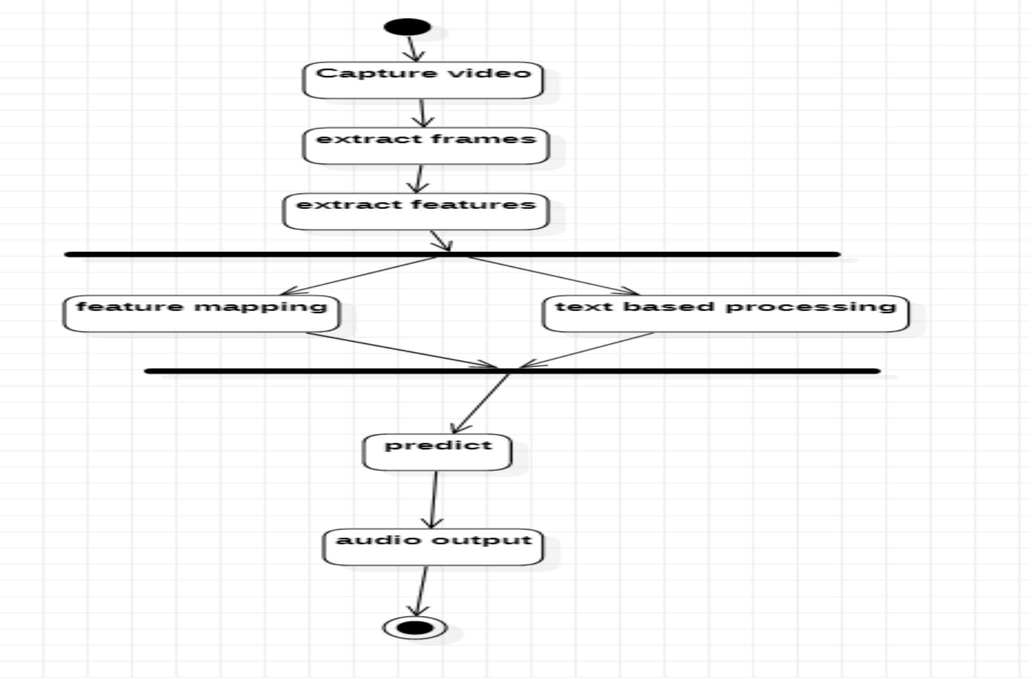


Fig 9: Activity diagram of video analysis model

7.3 Design Description

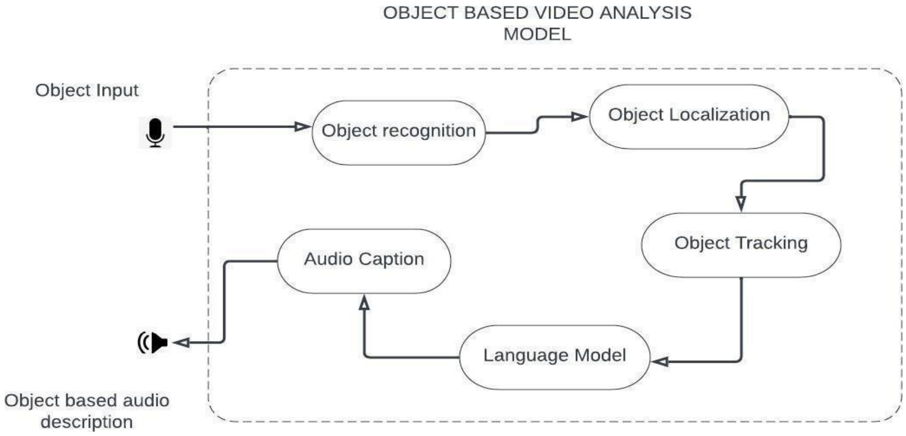


Fig 10: Flow diagram of Object recognition and classification

7.4 Swim lane diagram

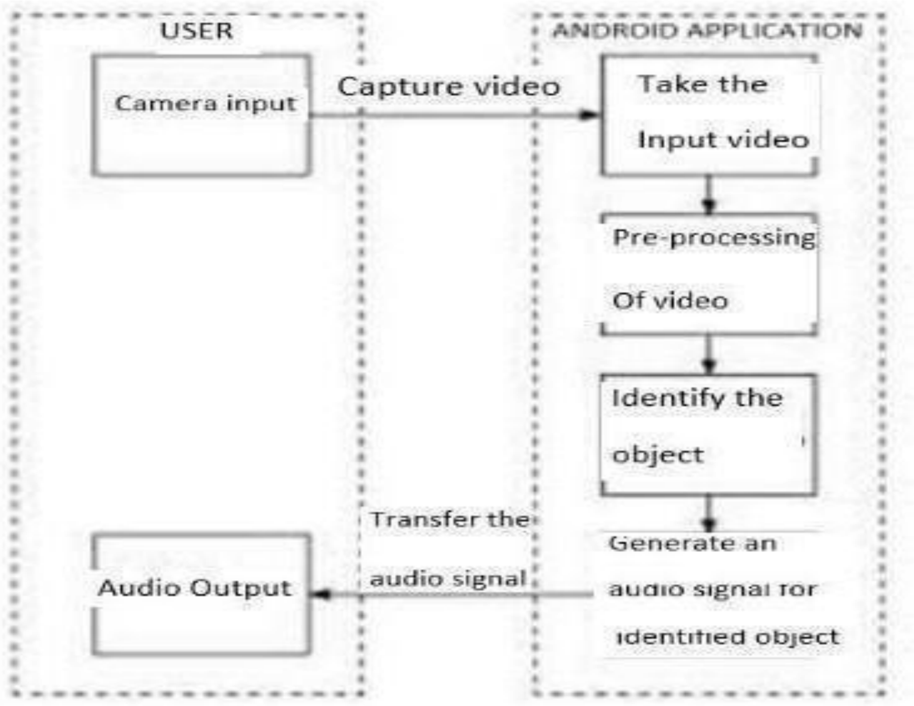


Fig 11: Swim lane diagram of data flow

7.5 User Interface Diagrams

The user has to login right when he/she installs the app. The user can login using their Google account as it is used in the later stage to save the data if the user wishes to. Once the user logs in, it takes him to the UI which is shown in the picture below. The recording is done for 5 secs and then processes it.

7.6 Packaging and Deployment Diagram.

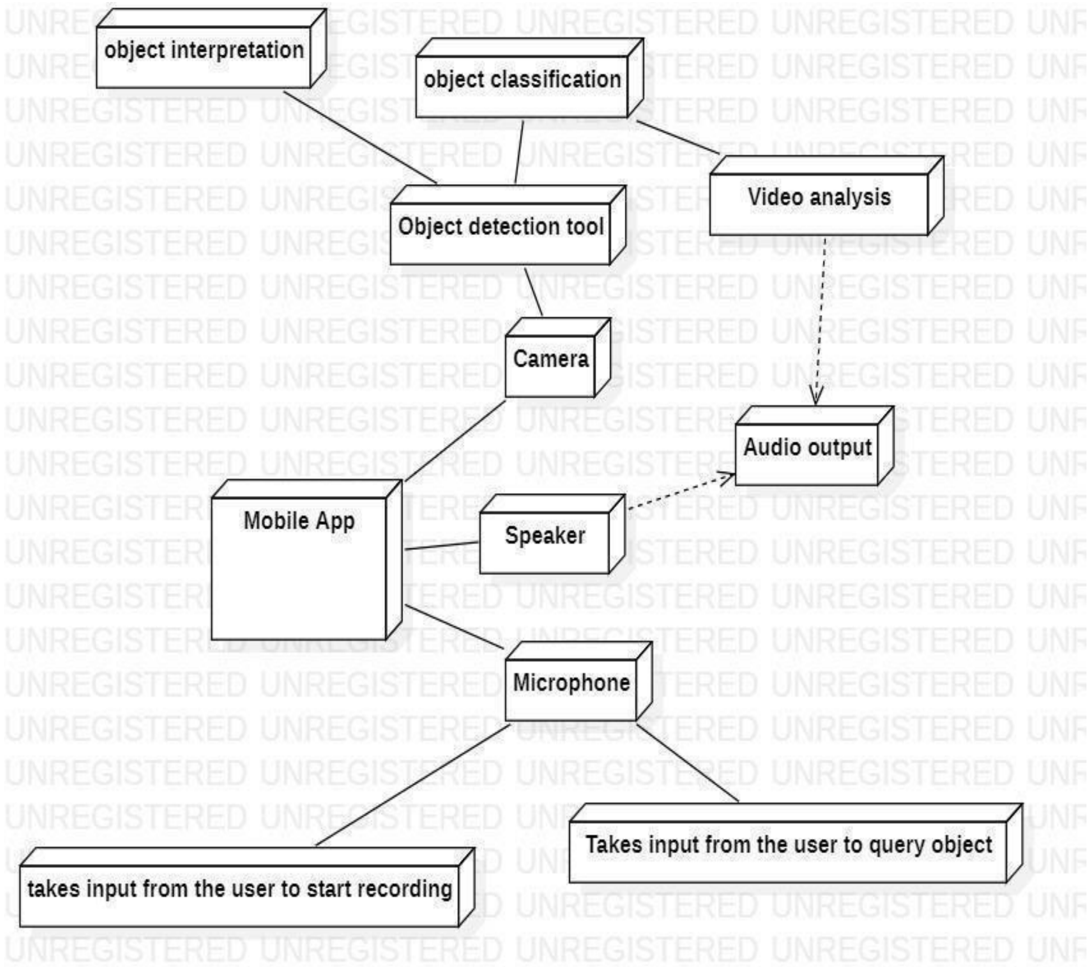


Fig 12 : Packaging and deployment diagram of the system.

7.7 Design Details

This application specifically is built to be used in an android platform. It must have basic hardware like a camera, minimum of 6GB RAM and a certain amount of internal storage for the app to run smoothly.

7.8 Novelty

We have a model that serves as a complete guide for a visually impaired person to perceive his surroundings to the best of his knowledge. There are features like object detection, video description, and distance estimation in a single model where every user gets the best. We also consider storage issues, so the video is nowhere to be saved in the device.

7.9 Innovativeness

Video analysis to generate one line description and distance estimation of objects with respect to the camera.

7.10 Interoperability

This application is designed to be run on an android platform.

7.11 Performance

The software's performance is reliant on the system's specification. The processing delay is minimal too.

7.12 Reliability

The software is very much reliable as it is almost hands free and the user can do most of the tasks without the help of any other person. Hence it can be reliable most of the times and keep the user safe and secure.

7.13 Portability

It is portable and can adapt to situations over time. It can be consistent with the upgrades and perform with the same efficiency.

7.14 Reusability

The object recognition based recognition models used here can be reused for different applications such as self-driving cars, and object tracking or in surveillance systems. Distance estimation could also serve as threat detection if channelised to a particular object.

7.15 Application compatibility

The application is compatible with the Android version. It is compatible with any android version and the UI would be the same for every device.

CHAPTER 8

PROPOSED METHODOLOGY

We have implemented a model in the domain of machine learning to help and aid visually impaired to understand their surroundings better. Considering multiple features and analyzing them deeply we have come up with models which work together better to serve a purpose. There can be multiple approaches to solve this problem, after brainstorming and many trial and errors later, we have curated the below approach. The input is taken from live camera of the smartphone and then the backend processes it.
The features that were to be implemented are:

- Object recognition
- Distance estimation if there are any humans nearby.
- Video analyzer and description generator for the recorded video.

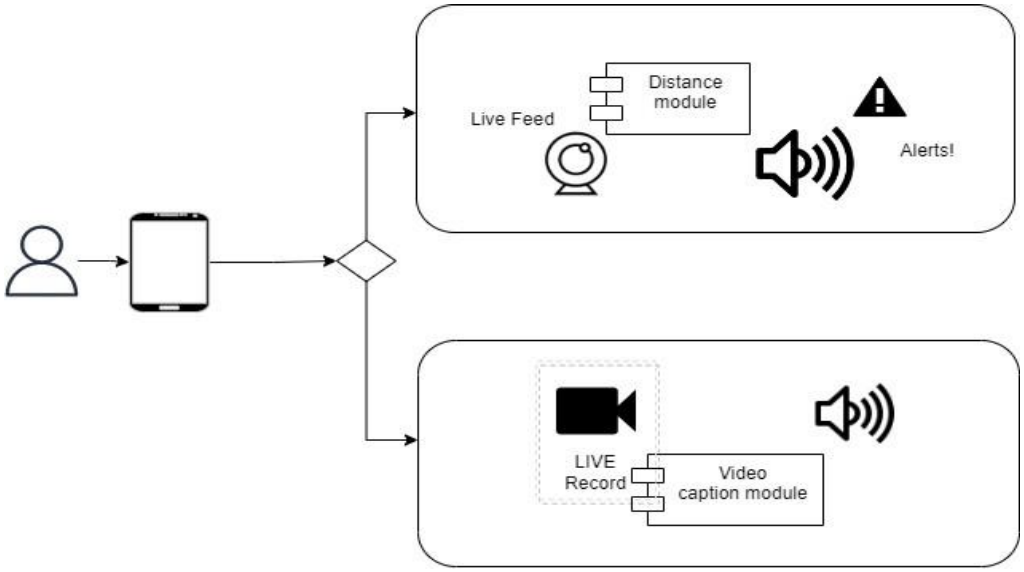


Fig 13: Architecture of Project

To implement the above features, we have used technologies that have shown us better results during trial and error.
Object Recognition: Since YOLO has better less localization errors it uses 224 X 224 images . It later tunes it to 448 x 448 for 10 epochs. Initially we started of with YOLO V3 , where frames were being extracted , but the accuracy and performance was a concern. Hence we proceeded with YOLO V5 model which is ‘You

only look once'. This module accepts the video input .It extracts and analyses the frames in that video. For every frame extracted , we have taken 3 frames per video for faster execution. Then YOLO algorithm draws bounding boxes for each frame and detected the objects present .

Yolo has a map of 94.63%. It has validation score of 0.8057. Yolo v5 has better accuracy than YOLO V4. We have a dataset which is sufficiently large and well labeled, we train it and establish a performance baseline. Generally YOLO v5 uses CNN to get features of the images. It has 80 classes. It is single stage object detector technique.

ARCHITECTURE OF YOLO V5 :

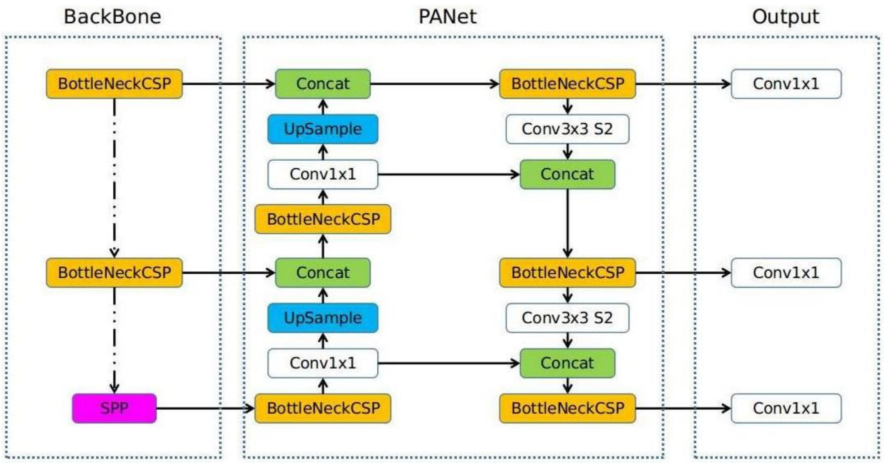


Fig 14: Architecture of Yolo v5 [17]

There are basically 3 layers :CSP Darknet, PANet and then the YOLO layer.

The darknet layer does all the feature extraction part, which is transferred to the next layer for feature fusion. The last YOLO layer outputs the detection results which is class, score , location and size.

Distance Estimation : We started with tensor flow for distance estimation. Here the objects were tracked and localized. But we have found a better approach where frozen graphs are used for object localisation.

Then we worked upon this to develop an inference graph which can accommodate multiple annotations. This file consists of features of the human face. Currently, our system is able to recognize human faces under this feature.

This module takes in live feed as input using a computer vision library. This live feed is fed as input to the analyser module, for every frame extracted we resize its dimension. We have used rgb (red, green, blue) since opencv loads images in bgr(blue green, red) which reduces the accuracy. There after, frozen graphs and inference graphs are used to train. We then map the faces detected to recognize it as humans. We have restricted the number of faces to be detected as 1 for now.

Then the distance is calculated using triangular similarity formula. We take a picture of our object using our camera and then measure the apparent width in pixels P. This allows us to derive the perceived focal length F of our camera:

$$F = (P \times D) / W$$

In Triangular similarity, the formula was derived using experimentation, to measure perceived focal length. Once this data is achieved, we use triangulation to calculate the distance of any human encountered. If a person is too close to the user, we provide a voice output alerting the user. If not, we activate the video description feature.

We have functions like ‘findcontours’ to identify the boundaries of an object in a frame. We are making the assumption that the contour with the largest area is our piece of paper.

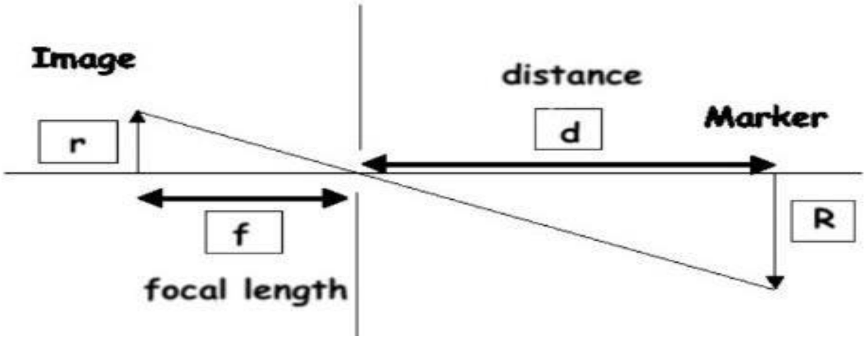


Fig 15:Triangular similarity estimation[18]

Based on the distance obtained, suitable decisions are made by the system. Estimations are completely audio output.

Video description analyzer:

The input is received from the live capture window from the android smartphone using a third-party application by specifying the http address of the webcam. The extracted frames from the live capture are given as an input to VGG16 for feature extraction. The features of each specified image is saved in the dictionary format and saved as a pickle extension file.

The LSTM model trained on a 30k image-caption dataset is then loaded from the directory, and the specific frame is passed as input to a function which retrieves the features of the specific image, the model is used to predict a index which is then converted to a word and then all the words is the end assembled with 'startseq' and 'endseq' to determine the starting and ending point of the sentence. After that, the predicted sentence is turned into audio by passing it through the pyttsx3 engine.

Architectures:

- VGG16

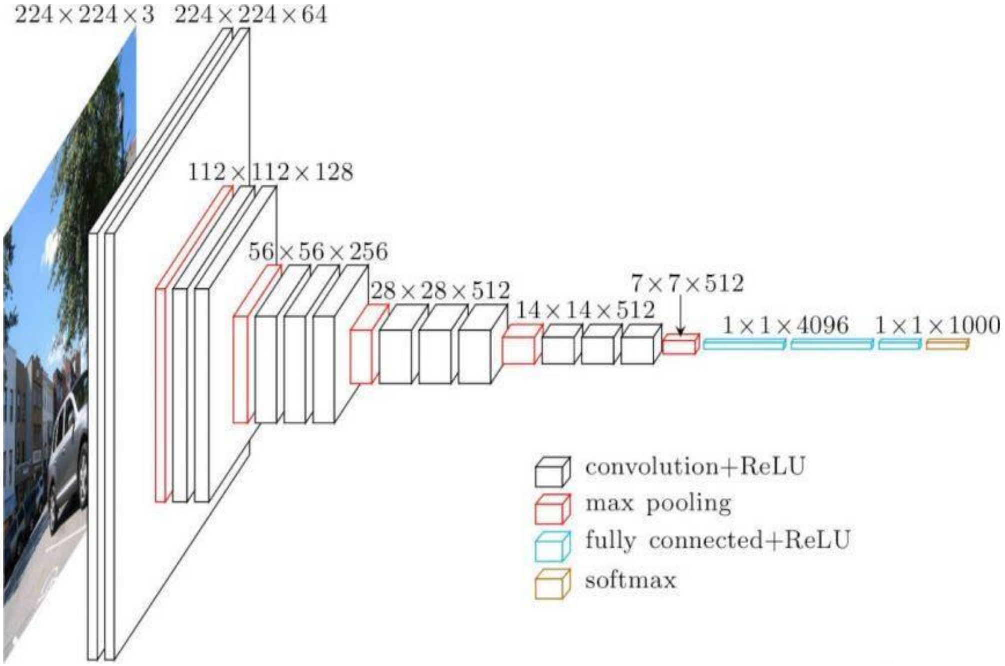


Fig 16: VGG 16 Architecture[19]

Layers:

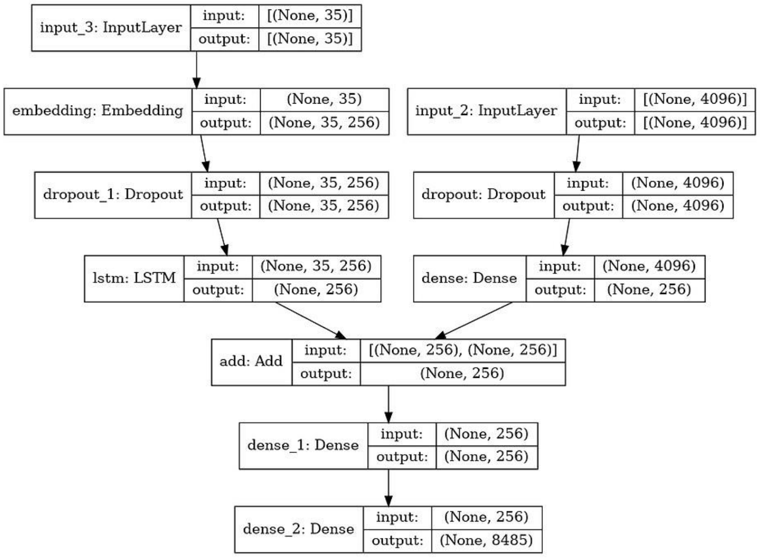

input_3: InputLayer | input: | [(None, 35)] | output: | [(None, 35)]
embedding: Embedding | input: | (None, 35) | output: | (None, 35, 256)
input_2: InputLayer | input: | [(None, 4096)] | output: | [(None, 4096)]
dropout_1: Dropout | input: | (None, 35, 256) | output: | (None, 35, 256)
dropout: Dropout | input: | (None, 4096) | output: | (None, 4096)
lstm: LSTM | input: | (None, 35, 256) | output: | (None, 256)
dense: Dense | input: | (None, 4096) | output: | (None, 256)
add: Add | input: | [(None, 256), (None, 256)] | output: | (None, 256)
dense_1: Dense | input: | (None, 256) | output: | (None, 256)
dense_2: Dense | input: | (None, 256) | output: | (None, 8485)

Fig 17: Layer distribution in VGG 16

- LSTM

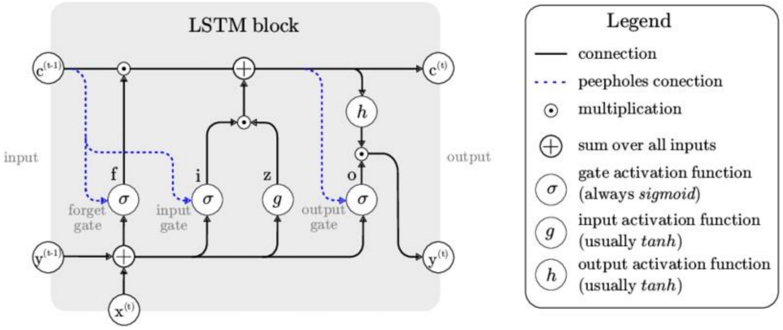


Fig 18: LSTM Architecture [20]

The caption generator uses two machine learning models to extract features and then another for text-based processing. The dataset used here is Flickr 30k dataset which has about 31000 images with 5 reference sentences each describing each image. For the image extraction, VGG (Visual Geometry Group)-16 is the first model used for feature extraction of the input image.

Then it's saved using the pickle library. Then we load the captions data and map the correct set of captions to the right image by assigning an “id” to each of them. Then preprocessing of the captions takes place and we split the data into training and testing sets. Now the second model that we use is the LSTM (Long short term memory) network for natural language processing.

CHAPTER 9

EXPERIMENTATION RESULTS AND DISCUSSION

9.1 Results obtained

Have a basic front end which is completely voice enabled. Login window, in case the user wishes to save his personal data. Scanning window which would be activated and set for 5 seconds. Video being captured and per frame extraction.

Our major expectancy was to identify the objects detected and label them accordingly. Video analysis and per frame detection to generate description accordingly. Video analysis and contour identification for the detected frames. Absolute distance estimation with respect to camera. There would be a Voice based output.

9.2 Discussion

We have a model that is able to accept inputs in the form of video and is restricted to a window. We wish to increase this processing speed and performance so that there are better results.

Our initial idea was to establish a model that would give an image description based on the object queried. Based on contrary views expressed in the discussion, we modified this feature to have a distance estimator that can serve as a threat detector for the user. We aim to establish a distance estimator that can estimate and establish contours for all kinds of objects. Currently, we have incorporated a system that establishes contours for human faces and analyzes distance with respect to the camer

CHAPTER 10

CONCLUSION AND FUTURE WORK

Concluding, we would say that there’s a vast scope for this project as our target audience are people with partial and full visual impairment. This app would be a feasible idea for the blind to rely on for their daily use. as it captures the vivid details and tries to describe them in the simplest manner.

Our model not only focuses on object detection aided with analysis of the video but also estimation of distance which could help the visually impaired people ward off potential danger and prevent mishaps, it allows them to see the world through the lenses of a camera making it easier for them to navigate through the currents of life. Since the model works on phones, there is no requirement for additional hardware components like a sensor, making it user friendly.

Our model involves live capturing and provides output with minimal delay which makes it even more lucrative and sustainable. It can also be used by detectives on a mission by enabling the response passage to a security network. In future, we would like to build upon the existing features and work on including features like SOS and emergency contacts, this in turn would help our target audience to improve their safety and security net.

REFERENCES

[1] B. Makav and V. Kılıç, "A New Image Captioning Approach for Visually Impaired People," 2019 11th International Conference on Electrical and Electronics Engineering (ELECO), 2019, pp. 945 -949, doi: 10.23919/ELECO47770.2019.8990630.

[2] A. Ali and M. A. Ali, "Blind navigation system for visually impaired using windowing-based mean on Microsoft Kinect camera," 2017 Fourth International Conference on Advances in Biomedical Engineering (ICABME), 2017, pp. 1-4, DOI: 10.1109/ICABME.2017.8167560.

[3] M. A. Khan Shishir, S. Rashid Fahim, F. M. Habib and T. Farah, "Eye Assistant : Using a mobile application to help the visually impaired," 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019, pp. 1-4, DOI: 10.1109/ICASERT.2019.8934448.

[4] T. M. Denizgez, O. Kamiloğlu, S. Kul and A. Sayar, "Guiding Visually Impaired People to Find an Object by Using Image to Speech over the Smart Phone Cameras," 2021 International Conference on Innovations in Intelligent Systems and Applications (INISTA), 2021, pp. 1-5, doi: 10.1109/INISTA52262.2021.9548122.

[5] David Bar-El, Thomas Large, Lydia Davison, and Marcelo Worsley. 2018. Tangicraft: A Multimodal Interface for Minecraft. In International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS). pp. 456–458.

[16] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6077–6086

[7] P. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077 - 6086, doi: 10.1109/CVPR.2018.00636.

[8] P. S. Rajendran, P. Krishnan and D. J. Aravindhar, "Design and Implementation of Voice Assisted Smart Glasses for Visually Impaired People Using Google Vision API," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1221- 1224, doi: 10.1109/ICECA49313.2020.9297553.

[9] Oneata, Dan & Stan, Adriana & Cucu, Horia. (2021). Speaker disentanglement in video-to-speech conversion.

[10] C. Rane, A. Lashkare, A. Karande and Y. S. Rao, "Image Captioning based Smart Navigation System for Visually Impaired," 2021 International Conference on Communication information and Computing Technology (ICCICT), 2021, pp. 1-5, doi: 10.1109/ICCICT50803.2021.9510102.

[11] S. M. Felix, S. Kumar and A. Veeramuthu, "A Smart Personal AI Assistant for Visually Impaired People," 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), 2018, pp. 1245-1250, doi: 10.1109/ICOEI.2018.8553750.

[12] L. G. Tornatzky and K. J. Klein, "Innovation characteristics and innovation adoption-implementation: A meta-analysis of findings," in IEEE Transactions on Engineering Management, vol. EM-29, no. 1, pp. 28-45, Feb. 1982, doi: 10.1109/TEM.1982.6447463.

[13] S. Firdus, W. F. W. Ahmad and J. B. Janier, "Development of Audio Video Describer using narration to visualize movie film for blind and visually impaired children," 2012 International Conference on Computer & Information Science (ICCIS), 2012, pp. 1068-1072, doi: 10.1109/ICCISci.2012.6297184.

[14] Description for Blind and Low Vision Users. In CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI ’21 Extended Abstracts), May8– 13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 7 pages.

[15] IBM Research, Tokyo Research Laboratory 1623-14 Shimo-tsuruma, Yamato, Kanagawa, 242-8502,Japan {mstm, kentarou, takagih, chie}@jp.ibm.com

[16] K. Miesenberger et al. (Eds.): ICCHP 2014, Part I, LNCS 8547, pp. 658–661, 2014. c SpringerInternational Publishing Switzerland 2014

[17] Katsamenis, Iason & Karolou, Eleni & Davradou, Agapi & Protopapadakis, Eftychios & Doulamis, Anastasios & Doulamis, Nikolaos & Kalogeras, Dimitris. (2022). TraCon: A novel dataset for real-time traffic cones detection using deep learning. 10.48550/arXiv.2205.11830.

[18] M. Pratama, W. Budi, S. A. Dimyani, A. Praptijanto, A. Nur and Y. Putrasari, "Performance of Inter-vehicular Distance Estimation: Pose from Orthography and Triangle Similarity," *2019 International Conference on Sustainable Energy Engineering and Application (ICSEEA)*, 2019, pp. 37-41, doi: 10.1109/ICSEEA47812.2019.8938648.

[19] J. Tao, Y. Gu, J. Sun, Y. Bie and H. Wang, "Research on vgg16 convolutional neural network feature classification algorithm based on Transfer Learning," *2021 2nd China International SAR Symposium (CISS)*, 2021, pp. 1-3, doi: 10.23919/CISS51089.2021.9652277.

[20] Yu Wang, "A new concept using LSTM Neural Networks for dynamic system identification," *2017 American Control Conference (ACC)*, 2017, pp. 5324-5329, doi: 10.23919/ACC.2017.7963782.

APPENDIX A DEFINITIONS, ACRONYMS AND ABBREVIATIONS

YOLO – You only look once algorithm

APP- Application