THIRD EYE

Capstone Project Report

END SEMESTER EVALUATION

Submitted by:

(101803613)	Utkarsh Sinha
(101803093)	Shubham Kumar Srivastava
(101803099)	Bineet Chadha

BE Third Year, COE

CPG No: 169

Under the Mentorship of

Dr. Anil Singh

Assistant Professor

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology, Patiala

July 2021

ABSTRACT

With the recent development of mobile processors, we now see more and more powerful smartphones in hands of people, processors inside these phones have more computing power than a normal desktop PC in 2000. This means a lot can be done on mobile devices that can make people's life easier. Especially nowadays mainstream mobile operating systems including iOS and Android had carefully thought about after accessibility functions built in. With these functions, visually impaired people can use a smartphone almost as conveniently as fair-sighted people. However, there is yet a need for some kind of system to be developed to help them "see" the world.

The first aspect of the project is to have a mechanism that runs on the phone and detects faces continuously. To make the most out of it in any environment, ideally, this function should run completely offline and should not require any cloud service. In that way, the user can use it when there is no network connection, and it also helps with minimising delays. In the short period of this project, our objective is to implement face detection, object detection, and real-time object search, all that using a single frame captured through the camera at the moment and then spell out the detailed description of the scene ahead. So the app would be developed in a way such that complex tasks can be performed while being lag-free and the user experience remains smooth. In the future, it can be expanded to be able to recognize other objects, identify people, gender classification, and read news.

Instead of using the traditional method of Eigenfaces and Fisher's faces, we would be using some new methodology that can work when the person in front is not properly facing the camera because that is the situation most often. We tried to explore the possibility of using a Deep Learning technique called Convolutional Neural Network (CNN) for our object detection and recognition purposes. Traditionally, distance of any object is computed using Ultrasonic sensors such as HC-sr04 or any other high frequency devices which generate sound waves to calculate the distance it traverses. But as we plan to work on a mobile application, we tried to make it more continental and feasible. In our project, the distance of the object from the camera is calculated using the depth information that the camera uses to draw the bounding boxes for localising objects.

DECLARATION

We hereby declare that the design principles and working prototype model of the project entitled Third Eye is an authentic record of our own work carried out in the Computer Science and Engineering Department, TIET, Patiala, under the guidance of Mr. Anil Singh during 6th semester (2020).

Date:

Roll No.	Name	Signature
101803613	Utkarsh Sinha	
101803093	Shubham Kumar Srivastava	
101803099	Bineet Chaddha	

Counter Signed By:

Faculty Mentor:

Dr. Anil Singh

Assistant Professor
CSED,

TIET, Patiala

ACKNOWLEDGEMENT

We would like to express our thanks to our mentor Mr. Anil Singh. He/She has been of great help in our venture, and an indispensable resource of technical knowledge. He/She is truly an amazing mentor to have.

We are also thankful to Dr. Maninder Singh, Head, Computer Science and Engineering Department, entire faculty and staff of Computer Science and Engineering Department, and also our friends who devoted their valuable time and helped us in all possible ways towards successful completion of this project. We thank all those who have contributed either directly or indirectly towards this project.

Lastly, we would also like to thank our families for their unyielding love and encouragement. They always wanted the best for us and we admire their determination and sacrifice.

Date:

Roll No.	Name	Signature
101803613	Utkarsh Sinha	
101803093	Shubham Kumar Srivastava	
101803099	Bineet Chaddha	

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Network
UI	User Interface
UX	User Experience
RPN	Region Proposal Network
API	Application Programming Interface
HMM	Hidden Markov Model
NLP	Natural Language Processing
LSTM	Long Short-term Memory
TTS	Text-to-Speech

INTRODUCTION

1.1 Project Overview

With the recent development of mobile processors, we now see more and more powerful smartphones in hands of people, processors inside these phones have more computing power than a normal desktop PC in 2000. This means a lot can be done on mobile devices that can make people's life easier. Especially nowadays mainstream mobile operating systems including iOS and Android had carefully thought about after accessibility functions built in. With these functions, visually impaired people can use a smartphone almost as conveniently as fair-sighted people. However, there is yet a need for some kind of system to be developed to help them "see" the world.

The first aspect of the project is to have a mechanism that runs on the phone and detects faces continuously. To make the most out of it in any environment, ideally, this function should run completely offline and should not require any cloud service. In that way, the user can use it when there is no network connection, and it also helps with minimising delays. In the short period of this project, our objective is to implement face detection, object detection, and real-time object-to-user distance measurement, all that using a single frame captured through the camera at the moment and then spell out the detailed description of the scene ahead. So the app would be developed in a way such that complex tasks can be performed while being lag-free and the user experience remains smooth. In future, it can be expanded to be able to recognize other objects, identify people, gender classification and reading news.

Instead of using the traditional method of Eigenfaces and Fisher's faces, we would be using some new methodology that can work when the person in front is not properly facing the camera because that is the situation most often. We tried to explore the possibility of using a Deep Learning technique called Convolutional Neural Network (CNN) for our object detection and recognition purposes. Traditionally, distance of any object is computed using Ultrasonic sensors such as HC-sr04 or any other high frequency devices which generate sound waves to calculate the distance it traverses. But as we plan to work on a mobile application, we tried to make it more continent and feasible. In our project, the distance of the object from the camera is calculated using the depth information that the camera uses to draw the bounding boxes for localising objects.

1.2 Need Analysis

Disability of any kind is a huge hindrance in one's way to using his/her physical potential to the fullest. As of late, much is being done by the government and the community as well to make the world more accessible to disabled persons. Parking lots are being modified to be accessible to the disabled ones, lifts and staircases are being designed in a way to help them commute easily through multi-floor buildings. Malls, theatres and other public places are being developed keeping in mind the adversities faced by the disabled humans. Despite all these efforts, there is still a need for a community more disabled-friendly.

According to the World Health Organisation, there are approximately 285 million people who have visual impairments, 39 million of them are blind and 246 million have a decrease in visual acuity. Almost 90% who are visually impaired are living in low income countries. In this context, Tunisia has identified 30,000 people with visual impairments; including 13.3% of them are blind. These Visual impairment present severe consequences on certain capabilities related to visual function:

- The daily living activities (that require a vision at a medium distance)
- Communication, reading, writing (which requires a vision closely and average distance)
- Evaluation of space and the displacement (which require a vision far)
- The pursuit of an activity requiring prolonged maintenance of visual attention.

Millions of people live in this world with incapacities of understanding the environment due to visual impairment. Although they can develop alternative approaches to deal with daily routines, they also suffer from certain navigation difficulties as well as social awkwardness. For example, it is very difficult for them to find a particular room in an unfamiliar environment. And blind and visually impaired people find it difficult to know whether a known person is talking to them or someone else during a conversation. Computer vision technologies, especially the deep Convolutional Neural network, have been rapidly developed in recent years. It is promising to use the state-of-art computer vision techniques to help people with vision loss.

All of these facts led us to the idea of developing a system which can easily be carried along while going outside and will be quite easy to use due to its simple user interface. Add to that the ability to verbally control the system and it will be even more convenient to use. Also, since the entire system will be in the form of a simple Android application, it has the potential to reach a large part of the population given the popularity and reach of Android-based smartphones. Considering the fact that our system is going to be an Android application, all it needs is just a few MBs of internet data and some storage space in the user's device to start using the system.

1.3 Problem Description and Scope

To make the world more convenient and easily accessible for the visually impaired persons, there is a need for some system which can verbally describe the current environment around the user. The aim of our project is to develop a voice-controlled, Android-based system that verbally describes the scenario ahead and also performs facial recognition while interacting with a human.

1.4 Assumptions and Constraints

TABLE 1: Assumptions and Constraints

S. No.	Assumptions
1	Good Network Connectivity & Bandwidth
2	Smartphone's camera is of sufficient resolution and image captured is of good enough clarity
3	Count of objects captured in the image is not very high and if human faces are captured then the no. faces is <=3.
S. No.	**Constraints**
1	Dataset used for training object detection model cannot cover every possible object label
2	The input video may not be very stable so it is possible to miss out on some objects in the scene

1.5 Approved Objectives

1. To make a mobile application that helps visually impaired people to see the world through using their verbal ability.
2. To provide a software that can detect people, objects, and their distance from the user and give a verbal description of the scenario ahead.
3. Moreover, it can distinguish between the known and unknown person in front of the user, if the person is known, it spells out his/her name and relation to that person through the headphones.

1.6 Methodology

Data collection

There are going to be more than 100 different classes for data like person, animals, car, bus, etc., which will be collected using the Google Dataset and Kaggle.

Labelling and pre-processing of data

The data will be cleaned, pre-processed and dimensionality will be changed as per our requirements, and at last, they are going to be labelled properly.

Model Training

Deep learning techniques and Machine learning will be used to train the model. There are going to be multiple ML models like object detection, virtual assistance, and image-to-text-to-speech.

Model testing

Testing will be done in parallel in order to have a minimum chance of error at every step of implementation. To check the accuracy of the trained model, various error techniques such as Root Mean Square Error, Mean Square Error, etc. will be used.

Android application

An Android application with easy-to-use UI and can be handled both verbally and by directly interacting with the device screen. The app runs completely offline on the mobile device and runs the ML-NLP models on the backend.

Application Testing

Testing will be done in parallel in order to have a minimum chance of error at every step of implementation. Both the Black box and White Box testing techniques will be used in order to have maximum accuracy.

1.7 Project Outcomes and Deliverables

1. A voice-controlled Android-based mobile application (virtual assistant) that helps visually impaired people to listen to things going around them through the headphones.
2. Two ML-NLP models, one, that can detect people, objects, and their distance from the user and second, which converts the image to text and then to audio signals which gives a verbal description of the entire scenario to the user through the headphones.
3. This application can distinguish between a known and unknown person. If the person is familiar to the user, it spells out the name and relation to that person, through headphones

1.8 Novelty of Work

Our novelty resides in an unified approach that deals with the interpretation of objects, scene, actions, and their mutual contextual constraints to improve action classification, scene context categorization, and semantic inferring. We believe that application goals will largely benefit from this perceptual framework.

The implementation of the Third Eye app is the sole part of the project, thought of and implemented by us. The techniques used by us for sub-problems of our project like Image Captioning is different from the usual approach of using direct algorithms. We are using the Transfer Learning technique to implement that module. Similar is the case for other modules.

REQUIREMENT ANALYSIS

2.1 Literature Survey

In recent years, motion related topics have been major concerns in the field of Computer Vision. Considering moving objects in real scenes, human beings are very important recognition targets, since face recognition and object recognition algorithms can greatly contribute to the realisation of automatic monitoring or guidance systems for various important applications. Recently, video surveillance and monitoring systems have gained importance because of security and safety concerns. Banks, borders, airports, stores and parking lots are the important application areas. As far as guidance systems are concerned, they find applications in areas like tourism, rescue missions, etc.

One of the important aspects of video monitoring and/or guidance systems is recognizing the scene ahead. There are two main parts in scenario recognition: Low level processing, including moving object detection, object tracking and feature extraction. Several new features have been developed in this level of processing, such as RUD (Relative Upper Density), RMD (Relative Middle Density) and RLD (Relative Lower Density), and other features such as Aspect Ratio, Width, Height, and Colour of the object are used along with these. High Level Processing includes event start-end point detection, activity detection for each frame and scenario recognition for sequence of images. This part is in the focus of research, and different pattern recognition and classification methods are being implemented and experimental results are being analysed daily so as to develop a robust technique.

To make a system accessible to people with some physical disabilities, it must be incorporated with features which provide hands-free and easy usage options. For a guidance system to be developed in such a way, it needs to generate a description of whatever video sequence is being fed into the system. Video description is the automatic generation of natural language sentences that describe the contents of a given video. It has applications in human-robot interaction, helping the visually impaired and video subtitling. Describing a short video in natural language is a trivial task for most people, but a very challenging one for machines. Automatic video description involves understanding of many entities and the detection of their occurrences in video employing computer vision techniques. These entities include background scenes, humans, objects, human actions, human-object interactions, human-human interactions, other events, and the order in which events occur. All this information must then be articulated using a comprehensible and grammatically correct text employing Natural Language Processing (NLP) techniques.

Another aspect of voice-controlled systems is voice recognition and speech-to-text conversion capability. Voice recognition techniques involve two broad steps – acquiring the audio signals through a microphone, and mapping of the acoustic signals onto textual sentences.

2.1.1 Theory Associated with Problem Area

Object detection and segmentation are vital use cases in numerous PC vision applications, for example, observation, vehicle route, and guided navigation. These technologies are being used widely and in varieties of applications. The idea of making use of such technologies for developing

assistant/guidance systems for people with disabilities has been under consideration for quite some time in the research community. Measures apart from technology are already being implemented in real life. Parking lots are being modified to be accessible to the disabled ones, lifts and staircases are being designed in a way to help them commute easily through multi-floor buildings. Malls, theatres and other public places are being developed keeping in mind the adversities faced by the disabled humans. Adding technology to all this can make it all more convenient and helpful. This exactly is what our project aims at.

2.1.2 Existing Systems and Technologies

Presently, certain products are available in the market in the form of mobile applications which are currently being used for vision loss or blindness.

1. Be My Eyes

The Be My Eyes app offers remote assistance from one of over a half million volunteer sighted helpers. The app uses your device's camera to create a video link, meaning you can point the camera and ask questions ranging from "Can you help me set my oven to 400 degrees?" to "My computer has stopped talking. Can you help me figure out why?" Be My Eyes began its life as an iOS only app, but recently it was also made available for the Android operating system.

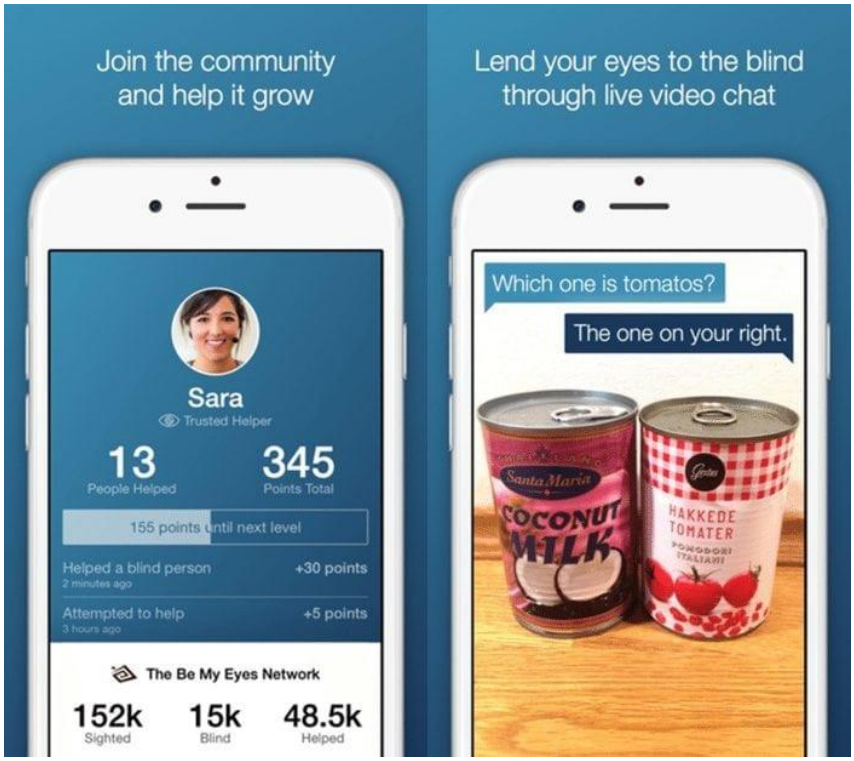

Figure 1: Existing Systems - Be My Eyes

2. Seeing AI

Seeing AI can recognize and speak text detected by the smartphone camera, either in tiny snatches or full pages at a time. It can read bar codes on grocery and other product labels, offer up the product name, and usually additional information, such as nutrition labelling, cooking and other instructions. Using Seeing AI you can snap pictures of your friends and family members, and later use the app to tell you who's nearby. An experimental setting can describe the scene around you, such as "A fenced-in yard," or "A blue door on an apartment building." You can also forward images you receive in email, or find on Facebook or Twitter, and Seeing AI will do its best to describe the action and read any text contained in the image.

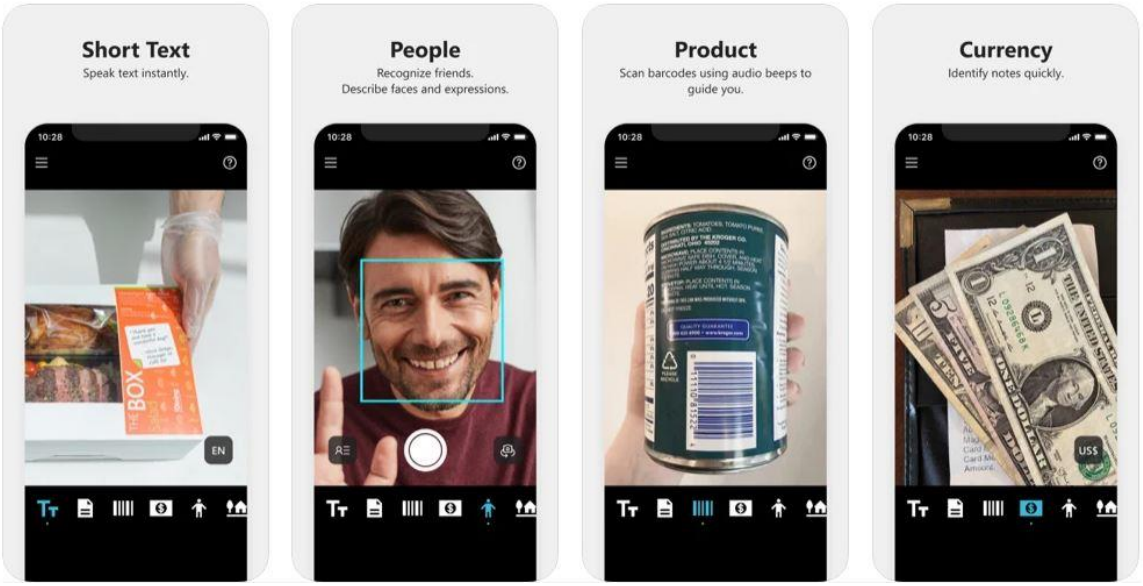

Figure 2: Existing Systems - Seeing AI

2.1.3 Research Findings for Existing Literature

First step in video monitoring/guidance systems is acquiring the video feed for further processing. Most of the current monitoring systems use multiple cameras and human operators to detect unexpected scenarios, because in realistic applications it is difficult to monitor a large target area at once, and difficult to track moving objects over a long time period. Considerable research has been performed on scenario recognition, including many approaches that can be distinguished in terms of:

- 2 dimensional or 3 dimensional
- single camera or multiple camera
- static camera or dynamic camera
- offline or online

Using one camera, Hongeng et al. [12] have recognized a sequence of several scenarios, called a multistate scenario, using Bayesian network and Hidden Markov Models (HMMs). Hamid [10] tracks objects with colour and shape-based particle filters, to extract features, and applies Dynamic

Bayesian Networks to recognize events. Yamato [8] describes a new human behaviour recognition algorithm based on HMMs; a sequence of frames is converted to a feature vector, and converted again to a symbol sequence by vector quantization. HMMs are trained, and the model that matches the event best is selected. Parameterized HMMs and coupled HMMs have been used to detect complex events such as the interaction between two moving objects. Existing approaches related to human action recognition include the top-down methods based on geometric body reconstruction and the bottom-up methods based on low-level image features.

To realise learning capability for time-sequential image data, Yamato employs the Hidden Markov Model (HMM), which can deal with time-sequential data and can provide time-scale invariability as well as learning capability for recognition. Although HMMs have been successfully used in speech recognition, HMMs have been applied to only a few problems in the computer vision field: planar shape classification, handwritten word recognition and modelling eye movement. In other words, HMMs have not been applied to motion recognition in general.

Amer [1] has worked on detection of events such as walking, sitting, and standing, using simple objects and clutter such as trees blowing in the wind and moving shadows. Davis [7] has worked on reliable recognition of basic activities from the smallest number of video frames. He used 2 probabilistic methods to detect simple activities such as walking, running and standing. Many other papers also address simple scenario recognition based on probabilistic methods.

Control charts approach is different from the above methods currently in use. It uses a rule-based system to categorise detected human activity into various classes, where the rules are obtained by control chart analysis. In essence, it treats the problem as analogous to controlling a manufacturing process. A process in control is analogous to the continuation of a sub-scenario and the time when the process goes out of control indicates that the tracked object is transitioning from one sub-scenario to another. A system for robustly tracking objects is used for the indoor surveillance application. Background subtraction has been implemented to detect foreground objects associated with regions that exhibit small changes over time. We adopt a luminance contrast method [11], [9] to reduce the side-effect of background subtraction, for two reasons:

1. It saves computational effort by using just one channel in colour images.

2. It removes much of the noise caused by luminance variations.

Video captioning research started with the classical template-based approaches in which Subject (S), Verb (V), and Object (O) are detected separately and then joined using a sentence template. These approaches are referred to as SVO-Triplets [2], [3]. However, the advent of deep learning and the tremendous advancements in CV and NLP have equally affected the area of video captioning. Hence, latest approaches follow deep learning-based architectures [4], [13] that encode the visual features with 2D/3D-CNN and use LSTM/GRU to learn the sequence.

Voice recognition techniques have also got their fair share of research. Kuldip K. Paliwal[6] and et al in the year 2004 had discussed that without being affected by their popularity for front end parameters in speech recognition, the cepstral coefficients which had been obtained from linear prediction analysis is sensitive to noise. Here, the use of spectral sub band centroids had been discussed by them for robust speech recognition.

Puneet Kaur et al [5] in the year 2012 had discussed how to use Hidden Markov Model in the process of recognition of speech. To develop an ASR (Automatic Speech Recognition) system the essential three steps necessary are pre-processing, feature Extraction and recognition and finally the Hidden Markov Model is used to get the desired result. Research persons are continuously trying to develop a perfect ASR system as there are already huge advancements in the field of digital signal processing but at the same time performance of the computer is not so high in this field in terms of speed of response and matching accuracy. The three different techniques used by research fellows are acoustic phonetic approach, pattern recognition approach and knowledge-based approach.

Other than these state-of-the-art techniques, the other option for implementing these features in a system is using APIs provided by various providers such as Google, Amazon, etc. Google's voice recognition and speech-to-text APIs are easily available for use in mobile applications. Object detection and facial recognition is available for use in mobile apps through Google's AutoML. Amazon's Alexa and Google Assistant are beautiful examples of good enough voice-controlled systems.

TABLE 2: Literature Survey

S. No.	Roll No.	Name	Paper Title	Tools/ Technology	Findings	Citation
1	Team member 1	Utkarsh	ARGMode-Activity recognition using graphical models	Dynamic Bayesian Networks	Object tracking, event recognition	Hamid [10]
2			Human action recognition using HMM with category separated vector quantization	Parameterized HMMs	Human Behaviour recognition	Yamato [8]
3			A reliable inference framework for recognition of human actions	Activity detection	Probabilistic approach with reduced frames	Davis [7]
4	Team member 2	Shubham	A computational framework for simultaneous real-time high level video representing	Hidden Markov Models	Motion Recognition	Amer [1]

5			Movie Description, Sequence to sequence-video to text	Deep learning (CNN+LSTM)	Video/Image captioning	Schiele [4], Saenko [13]
6			Speech Recognition with Hidden Markov Model: A Review	Automatic speech recognition	Using HMM in recognition of speech	Puneet Kaur [5]
7	Team member 3	Bineet	People tracking in surveillance applications	Human activity detection	Control Charts approach	Velastin[9], Kanade [11]
8			Natural language description of human activities from video images based on concept hierarchy of actions	Video captioning	SVO-triplets approach	Salvi [2], Fukunaga [3]
9			Recognition of Noisy Speech using Dynamic Spectral Subband Centroids	Voice recognition	Use of spectral subband centroids	Paliwal [6]

2.1.4 Problem Identified

The main problem with Object Detection and Recognition technology in the current scenario is that objects may not get recognized and some objects in an application are not recorded in the same way that other objects from the same domain are recorded. There may be a difference in lighting or rate of motion or the position due to which the object may not be properly recognised. Another issue is that even if an object is getting detected it may be classified differently as compared to previous classes.

Many tools available in popular programming languages for Object Recognition are still computationally intensive to be compatible with real time systems. Real time systems have the requirement of processing video frames in quick succession as they are generated very fast in live cameras.

2.1.5 Survey of Tools & Technology

We are using multiple machine learning technologies for this project. Like, Object detection, Face detection, facial recognition, Image captioning, Text-to-speech, Audio-to-text. For object detection we are using YOLO-Tensorflow ML-model as it is very fast and accurate. For facial detection and recognition we are using the Haar-Cascade facial feature detector with KNN algorithm. For Image captioning we are using Neural Network algorithms and transfer learning techniques. For Text-to-speech and Audio-to-text we are using Google APIs and Libraries.

These strategies are used by us in python in the strategy video-catch. Further utilisation of library imutils is made to control pictures utilising gaussian haze, and so on. We have partitioned the video into parts utilising argparse video altering library.

2.2 Software Requirement Specification

2.2.1 Introduction

The basic outline of the working of the project is that using a smartphone's camera, a live video is recorded, processed and objects (and faces, if any) are detected in the captured video frames. The detected information is then transformed into natural language and verbally informed to the user. The system is in the form of a mobile application and keeping in mind that it is aimed at servicing blind users, it will have voice-control apart from the usual touch control.

2.2.2 Purpose

In today's scenario, visual impairment is a major issue. The part of the human population suffering from partial or complete blindness has been on a rising curve since the past few years. These people struggle performing their daily chores and roaming outside their homes is like a haunting challenge for them. They require someone or something to guide them to their destination and to assist them in their tasks. The basic concept behind this project is to develop a system which will be readily available and will assist the blind user in his/her tasks besides guiding them on their way outside their house. Such a system built with good enough accuracy can be very handy for such use cases and making it available in the form of a mobile application makes sure it reaches a large part of the population.

2.2.3 Intended Audience and Reading Suggestions

This document follows IEEE specified format for writing Software Requirements Specification. The system features specified further in this document are arranged in order of their priorities and are described in paragraphs. The intended audience for this document are:

- Developers
- Designers
- Testing team members

2.2.4 Project Scope

This project is mainly aimed at providing guidance to people with visual illness. But apart from this

use case, it can also be handy in scenarios like video surveillance, verbal broadcast of events, etc. The system is capable of detecting and recognising human faces and can thus be used in criminal investigations. It also has capabilities to search for a particular object in the input image. This feature can have household applications or even in legal search investigations. The ability to verbally describe live camera feed can also be exploited to monitor a defined space for unusual activities.

2.2.4 Overall Description

2.2.4.1 Product Perspective

The main idea behind this project is to provide some sort of assistance to people with visual impairment. People with acute visual illness face many difficulties performing their daily chores. For any kind of movement outside their home, they have to rely on someone to guide them on the streets. Even for finding a particular object, either they have to touch it (which can be dangerous, for ex. knife) or they have to ask someone to help them out. This app aims to act as an always available assistant to the user so that they don't have to rely on anyone else.

2.2.4.2 Product Features

1) Stand-alone mobile application based on Android OS.
2) Voice-controlled handling.
3) Verbal commentary of live camera feed.
4) Recognition of known human faces.
5) Searching for a particular object in the scene ahead

2.2.5 External Interface Requirements

2.2.5.1 User Interfaces

The app supports voice-controlled handling which needs to be initiated by the user. The speech recognition module becomes active on detecting any vocal input and records and processes the input audio to decode user commands. Accordingly, a suitable response is generated and is outputted to the user. The user has option to perform following actions:

- ➢ Give voice inputs to open or close the app
- ➢ Switch to object search mode or known face entry
- ➢ Raise any query (through voice) about some feature in the app

2.2.5.2 Hardware Interfaces

- ➢ Android Smartphone (with rear camera)
- ➢ Cloud GPU

2.2.5.3 Software Interfaces

- ➢ Firebase Realtime Database
- ➢ Google TTS API
- ➢ ML Kit - Google Cloud Platform
- ➢ Tensorflow

2.2.6 Other Non-functional Requirements

2.2.6.1 Performance Requirements

The foremost priority of our project is to process and detect objects in the live scene as quickly as possible. The latency in this entire process should be below acceptable limits. Apart from that, reasonably high accuracy of the object detection module is crucial for the system. Database access and modification for known faces should be possible in the blink of an eye. Distance calculation from captured images is based on relatively new technology and thus cannot be expected to be near accurate but we aim to develop it accurate enough to correctly determine at least the range of the obstacle location.

2.2.6.2 Safety Requirements

The most important part of the project that requires validation before access is the database used to store known face details. The database does not provide read/modify access to the users. Users can only add data to the DB. Data is fetched from it for recognition purposes by the ML models that run on a separate platform and are like a blackbox for the users of the app. In case of accidental shutdown, the user is notified through an audio response. Obstacle detection is a crucial part of the system due to the fact that the user's safety is involved and so obstacle alert is issued at some distance less than the threshold to account for the error in calculation.

2.2.6.3 Security Requirements

The only type of data that is being stored in our case is that of the known faces and that being stored on Firebase is upon the security of Google Cloud platform. The database is protected against unauthorised write/modify access. The live feed recorded for verbal commentary to the user is not stored in any form eliminating any possibility of its misuse. Faces captured for storing into the database are stored in the form of feature vectors and not the face image itself to further avoid any misuse in case of data leak.

2.3 Cost Analysis

Since our project does not require any external hardware, it is difficult to calculate the exact cost incurred. We can work out an estimate of the total cost based on the effort put into the project and the basic resources required like electricity, internet, etc. Using the following formula for calculating

effort-months (assuming total schedule to be 6 months),

$$Schedule\ (\ in\ months) \ = 3 \times \ (\ eff\ ort - months)^{\frac{1}{3}}$$

It comes out to be 8. Now if we price 1 effort-month at Rs 100, the total cost amounts to Rs 800.

2.4 Risk Analysis

The table shows the risks involved in the project along with their probabilities of occurring and the impact if they do occur.

TABLE 3: Risk Analysis

S. No.	Risk	Probability	Impact
1	Application shuts down abruptly	0.1	9
2	Camera switches off	0.3	9
3	Some objects are missed out in the scene	0.6	6
4	A known face is termed as unknown	0.2	5
5	Obstacle distance is wrongly determined	0.3	8

METHODOLOGY

3.1 Investigative Techniques

TABLE 4 : Investigative Techniques

S. No.	Investigative Projects Techniques	Investigative Description Techniques	Investigative Projects Examples
1	Experimental	For the image captioning module, we have used a transfer learning approach. We developed a pipeline where the output from CNN is passed on to a LSTM network and then word embeddings are created. From this, image captions are generated.	Fast-CNN for object detection, RNNs for sequence determination projects, etc.

3.2 Proposed Solution

With the aim of assisting blind people, the system we propose is an Android based mobile application that can be controlled through voice commands. It uses the smartphone's camera to capture live scenes ahead of the user and then detects different objects in the scene. Along with the objects, if any human face is detected it is sent for further processing where it is searched for in the known faces database. If the face is already present, that person's name and relation to the user is spelled out. After all this processing and detection tasks, the entire information is compiled together and is transformed into meaningful textual format. This text information is converted into natural language (if not already) and using Text-to-speech functionality a verbal commentary of the scenario ahead is provided to the blind user. Apart from these primary features, the app also provides the user the functionality to give voice commands and the app will respond accordingly. The user can even search for a particular object in the live scene.

3.3 Work Breakdown Structure

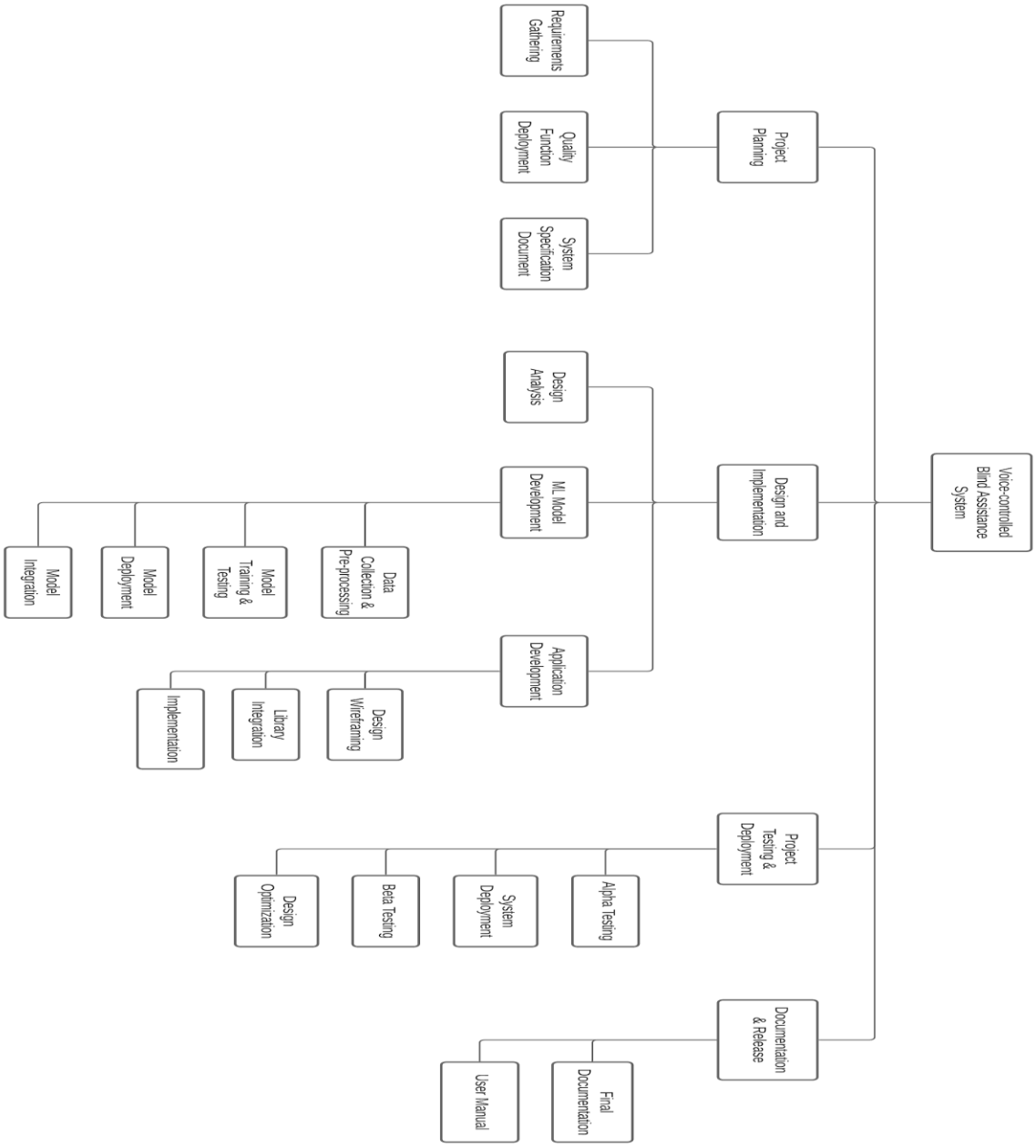

Figure 3 : Work Breakdown Structure

3.4 Tools and Technology

- ★ Android Studio
- ★ Jupyter Notebook
- ★ Tensorflow
- ★ Firebase
- ★ Google Speech API
- ★ GCP ML Kit

DESIGN SPECIFICATION

4.1 System Architecture

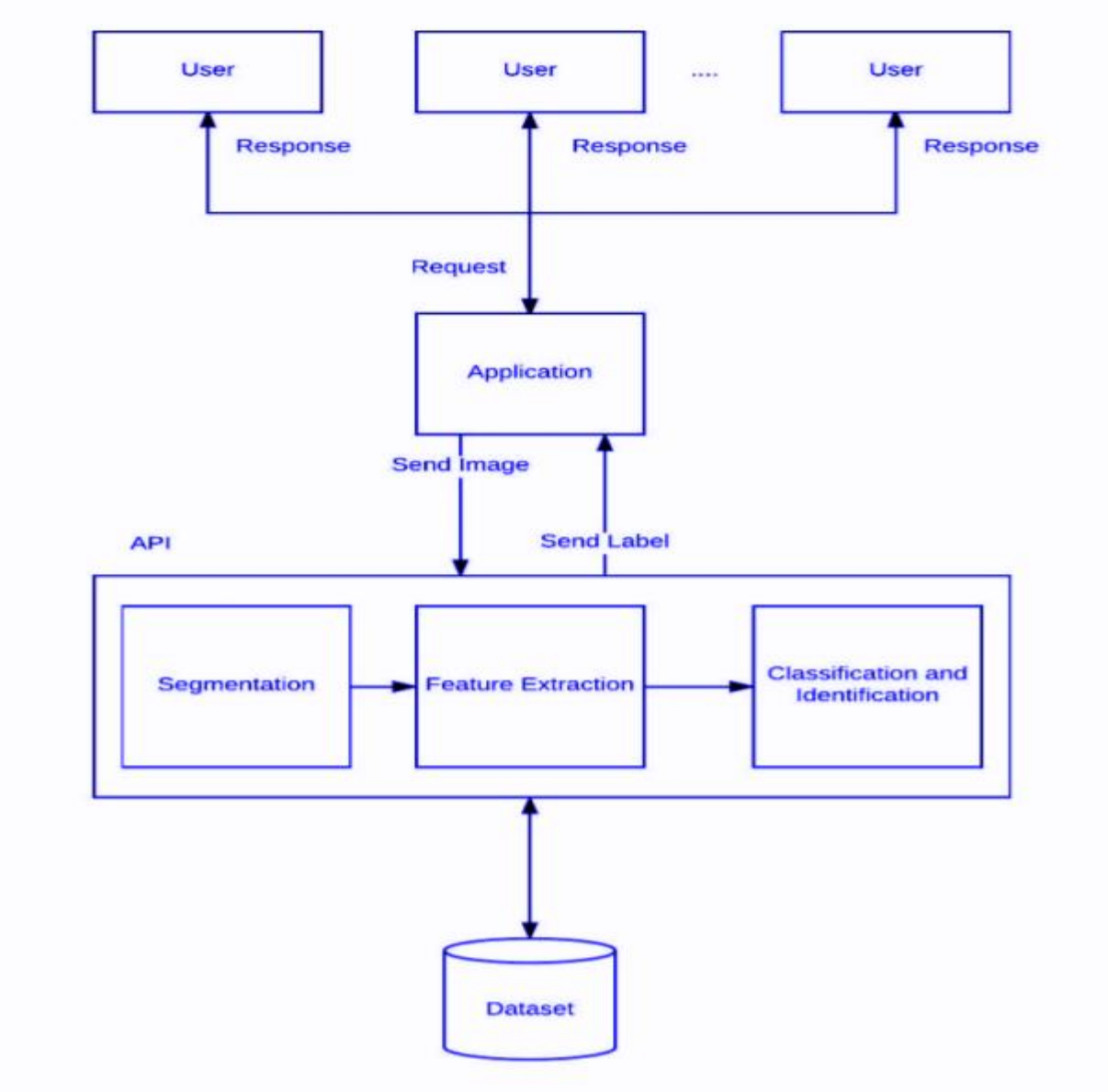

Figure 4 : Architecture Diagram

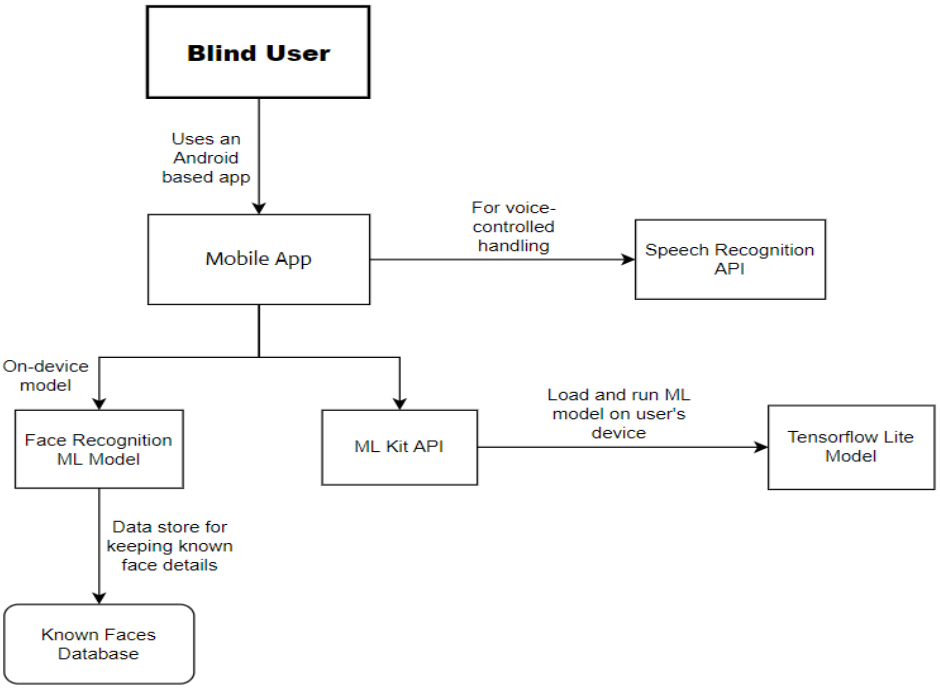

Figure 5 : System Design Overview

4.2 Design Level Diagrams

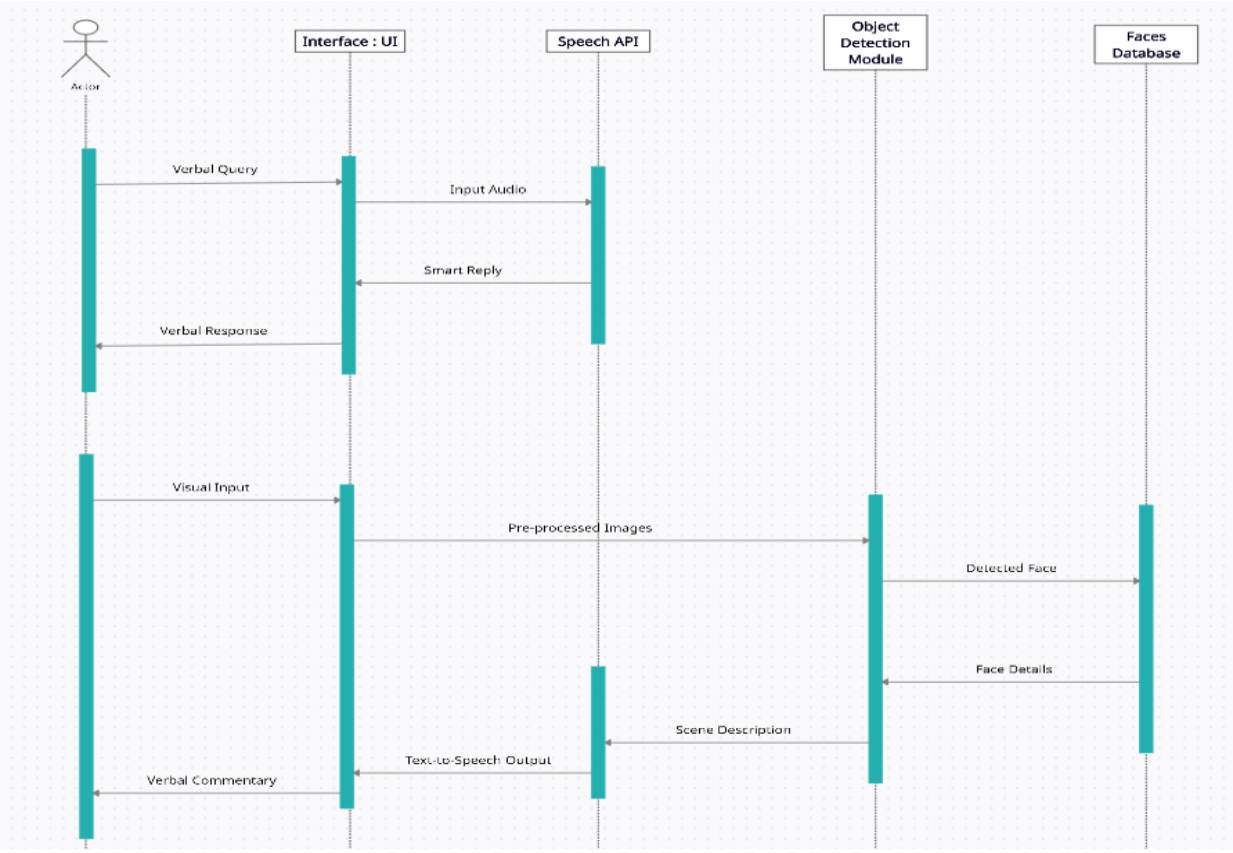

Figure 6: Sequence Diagram

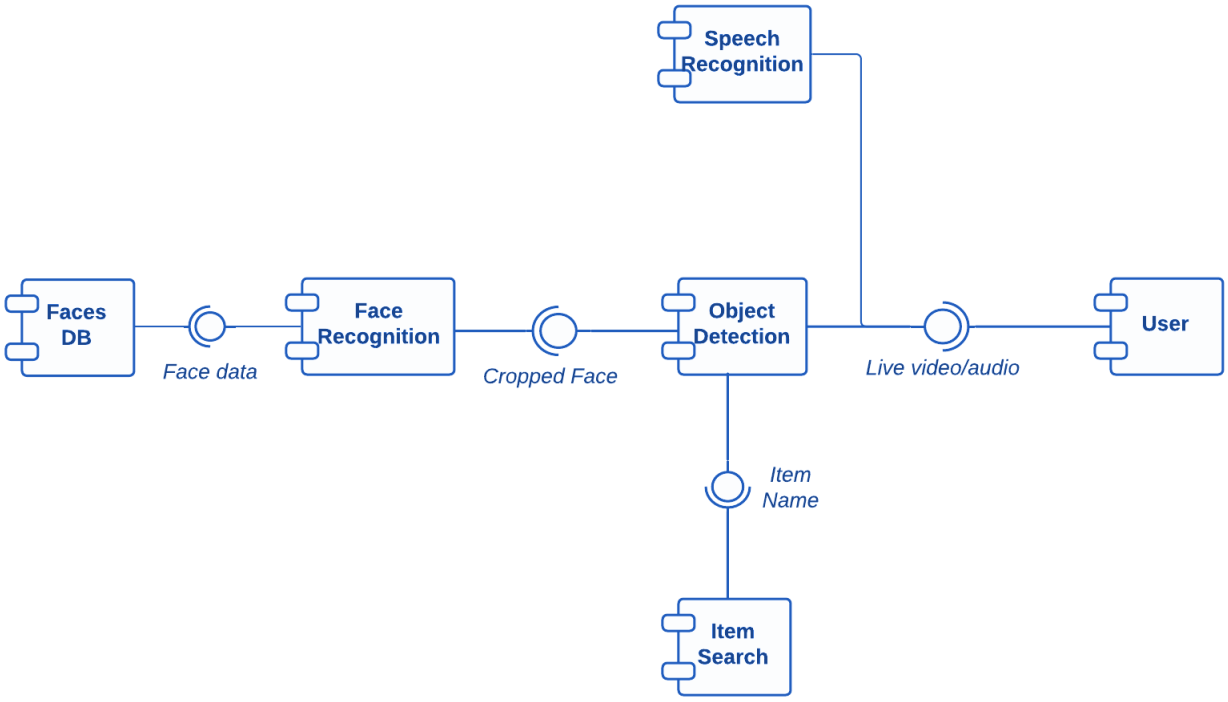

Figure 7: Component Diagram

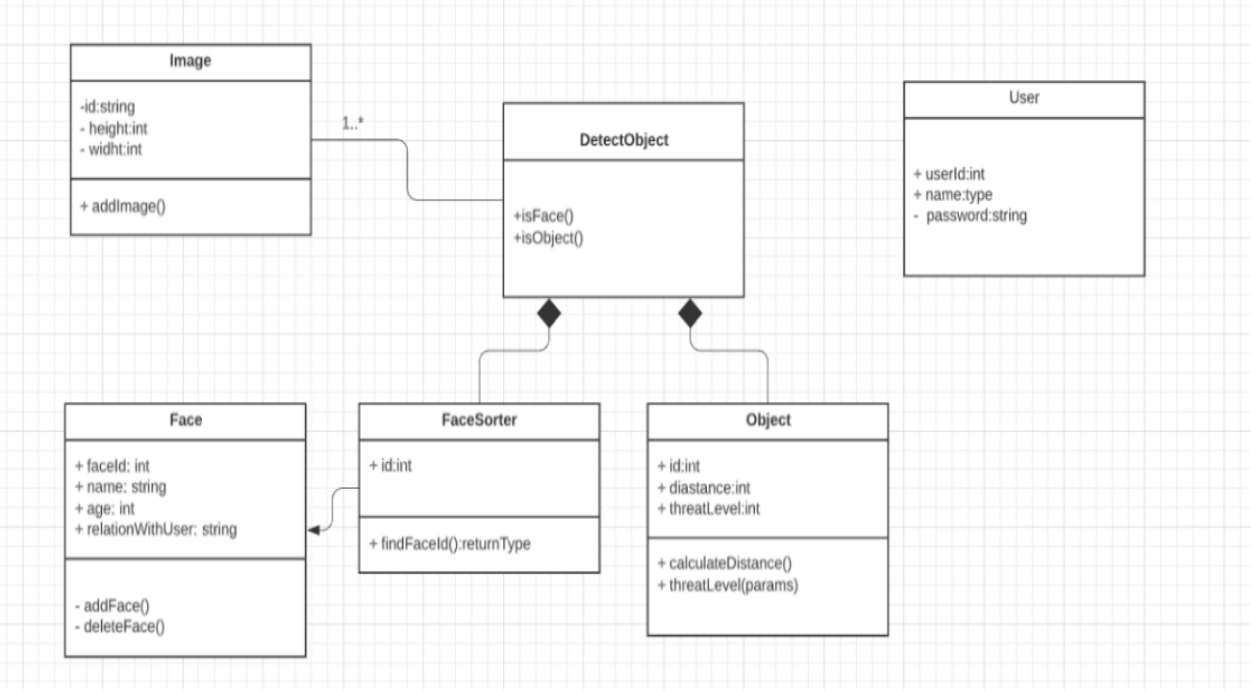

Figure 8: Class Diagram

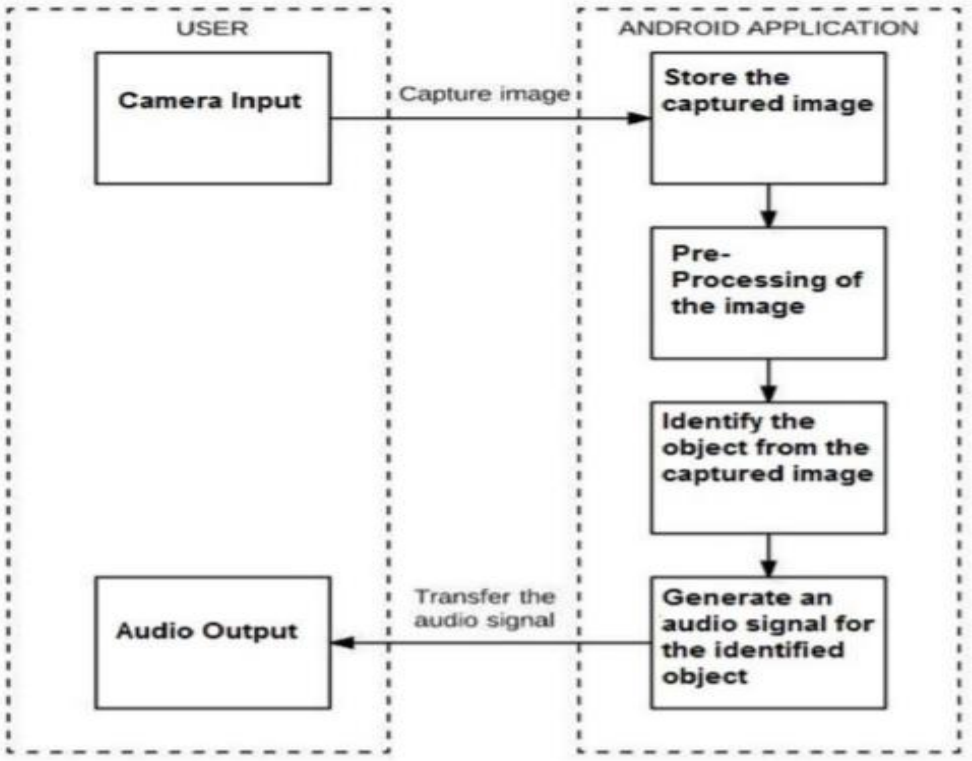

Figure 9: Module Diagram

4.3 User Interface Diagrams

Figure 10: GUI Design

IMPLEMENTATION AND EXPERIMENTAL RESULTS

5.1 Experimental Setup

Third Eye is also one of the technological enhancements, whose aim is to make the world more convenient and easily accessible for the visually impaired persons, there is a need for some system which can verbally narrate the current environment around the user. The aim of our project is to develop a voice-controlled, Android-based system that verbally describes the scenario ahead, helps find particular objects and also performs facial recognition while interacting with a human.

Taking the scenario when the person is alone and want to find a way to go upstairs, or through the door, and go to the washroom, bathroom, or somewhere on the street, in garden, this application will help him/her to find those ways and guide him/her properly throughout his/her way. Or let's say the user wants to find his spectacles, water bottle, and other basic things, this app will help him/her to find those objects. The user just has to give the command to the application assistant and the application will help him find that object. Or let's say the user has met a person and if the user wants to know who that person is and what the relation is, is this person known or unknown? Here we are making this application that can detect people and objects in proximity of the user and give a verbal description of the scenario ahead. It can distinguish between the known and unknown person in front of the user, if the person is known, it spells out his/her name and relation to that person through the headphones. We have developed an algorithm for detecting these things, image frames are captured regularly and processed on the spot and are not saved to hard disk or on the cloud. The details like known person name, their facial data are stored in the hard disk and backed-up to the cloud.

5.2 Experimental Analysis

5.2.1 Data

1. For the initial phase of the testing, we used already posted image data from the internet, which included videos of cars, buses, bikes, people, birds, cats, cows, dogs, horses, sheeps and other animals, and traffic signals.

2. After getting appropriate results the data was now the images captured from the application of known persons.

These data values also provided results up to the expectations.

5.2.2 Performance Parameters

We evaluated our Third Eye application on different parameters.
Average app start-up time → 800ms
Facenet model for facial recognition, inference time → 390ms
Machine Learning Model accuracy → 69.72%
Frame rate of live camera feed → 28fps
Frozen frames on Explore screen → 12.25%

5.3 Working of the project

5.3.1 Procedural Workflow

The working of the entire system is depicted by the flowchart below.

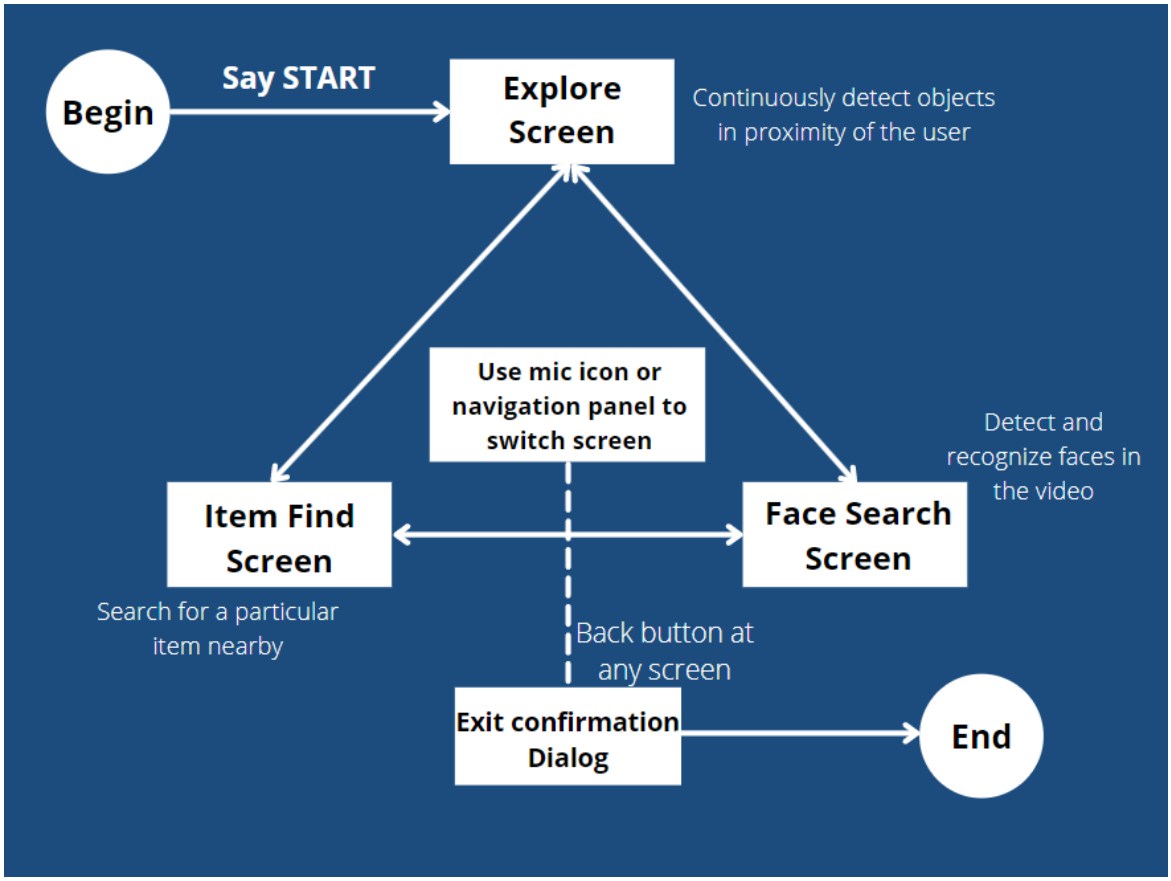

Figure 11: Procedural Workflow

The system starts when the app icon is clicked on or when the user uses Google assistant to open the app. Once started, the user needs to speak 'START' as this is the keyword used to kickstart the whole thing. The next screen that the user navigates to is the Explore screen. This module of the app is responsible for continuously detecting and speaking out the objects in close proximity to the user. It currently supports about 400 labels ranging from daily use items to animals to public items. The technology used behind the scenes is the famous CNN model.

At this stage, the user now has two more options to navigate to - face recognition space and Item find screen. To switch to these screens, one can either use the navigation bar at the bottom of the screen or since this app has voice-support, using the mic button at top-right, can activate the speech module for navigation.

Let's first analyse the Face recognition feature. After moving to this screen, the system continuously searches for human faces. Once found it generates face embedding of that face and then compares it with all available faces in the database using cosine or L2 similarity. The one which best matches is presented as the output. Following this the name and relation of that person with respect to the

user is spoken out. If the detected face does not match any available face, then it is termed as Unknown and the user is prompted to enter the name and relation of this new person which is then saved into the database.

Next screen is the Item Find space. Here the user is presented with a list of some common items. From this list an item is selected and then the system searches the user's proximity for that particular object only. Once found, the user is informed verbally, visually and even through haptic feedback by vibrating the smartphone.

Exiting the application is pretty straightforward, tapping the back button at any screen pops a confirmation dialog. As per the interaction with the dialog, the user either remains on the app or exits it.

5.3.2 Algorithmic approaches used

For Face detection and recognition, we have used the FaceNet approach. FaceNet is a Google-developed facial recognition system that obtained state-of-the-art results on a variety of face identification benchmark datasets in 2015. (99.63 percent on the LFW). The concept of triplet loss was introduced in that paper. For implementation on mobile, a variation of this approach is used which resulted in a model file size of just 4.0MB which after conversion to tflite model for Android was of size 5.2MB.

CLASSIFIER	ACCURACY	PRECISION	RECALL	ROC
SVM	85.68%	0.86	0.87	0.86
Decision Trees	84.61%	0.85	0.84	0.82
KNN	86.32%	0.86	0.86	0.88
ANN(for 100 epochs)	83.10%	0.88	0.87	0.88
CNN(for 300 epochs)	91.11%	0.93	0.89	0.97

Figure 12: Accuracies of different ML Models

For object detection and labelling, CNN model was used. The workflow of the algorithm is as follows:

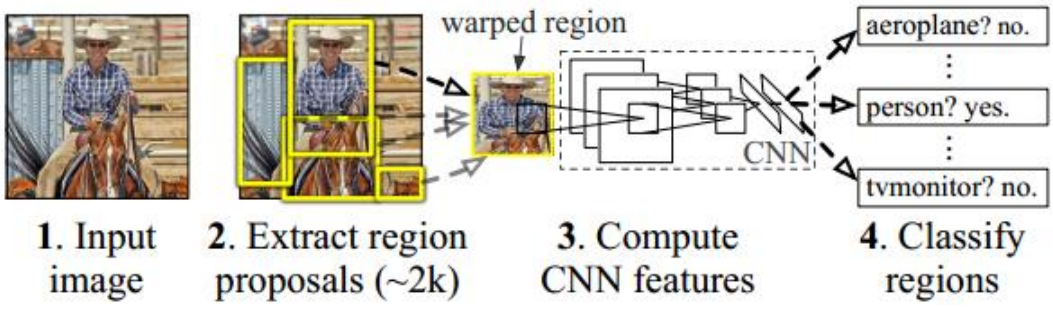

Figure 13: CNN algorithm steps

The algorithm variants currently in use are Fast-CNN, Mask-R CNN, etc. These are significantly faster and more accurate than the original neural network model.

For speech recognition and verbal correspondence with the user, we have used Google's SpeechRecognition API. Behind the scenes it makes use of LSTM RNNs which is a major improvement over the traditional Gaussian Mixture Model (GMM) technique. Android's SpeechRecognizer class provides an easy and efficient way to implement this functionality and that's what we have made use of.

1: **algorithm** Parallel-CNN
2: **input**: d: dataset, l: dataset true labels, W: Word2Vec matrix
3: **output**: score of Parallel-CNN trained model on test dataset
4: **let** f be the featureset 3d matrix
5: **for** i in dataset **do**
6: **let** f_i be the featureset matrix of sample i
7: **for** j **in** i **do**
8: $v_j \leftarrow vectorize_{(j, W)}$
9: **append** v_j to f_i
10: **append** f_i to f
11: f_{train}, f_{test}, l_{train}, $l_{test} \leftarrow$ split feature set and labels into train subset and test subset
12: M ← Parallel-CNN (f_{train}, l_{train})
13: score ← evaluate (i, l_{test}, M)
14: **return** score

Algorithm 1: Pseudo code of CNN

5.3.3 Project Deployment

The different components of the system can be easily understood by the component diagram below:

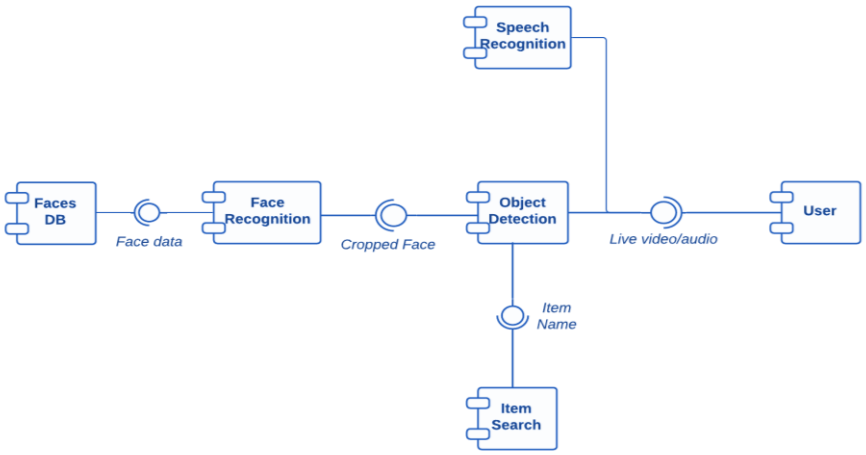

Figure 15: Project Components

The project has four major components - Object detection, Item Search, Facial recognition and Voice support. The user realises the interface exposed by the entry component which is the object detection one. Based on the live camera feed, it either processes the frame or extends the partially-processed frame to either of other two related components - Item Search and Facial recognition. The database for storing data of known faces is stored locally on the user's device and is backed up by its cloud-based copy. The voice control system makes use of the device's microphone and speakers whereas the intermediate processing is done through Google's API for this purpose.

The actual setup of the system can be understood by the deployment diagram below:

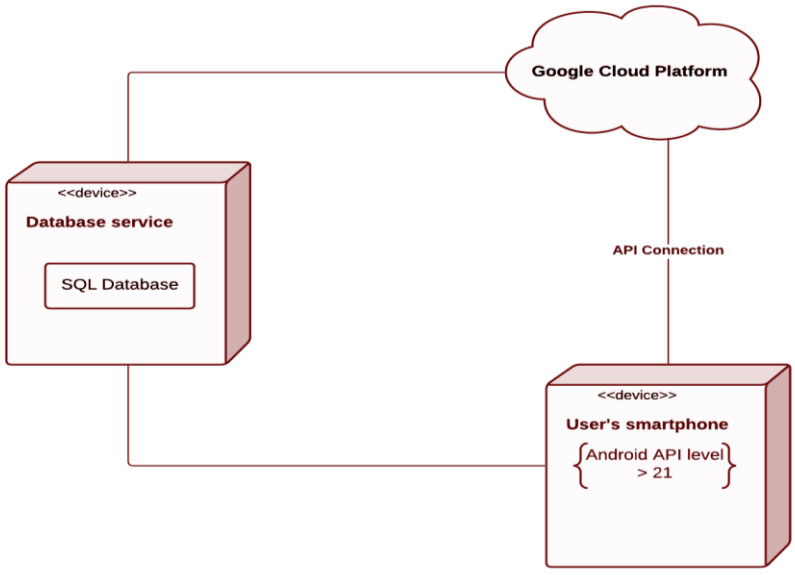

Figure 16: Deployment diagram

5.3.4 System Screenshots

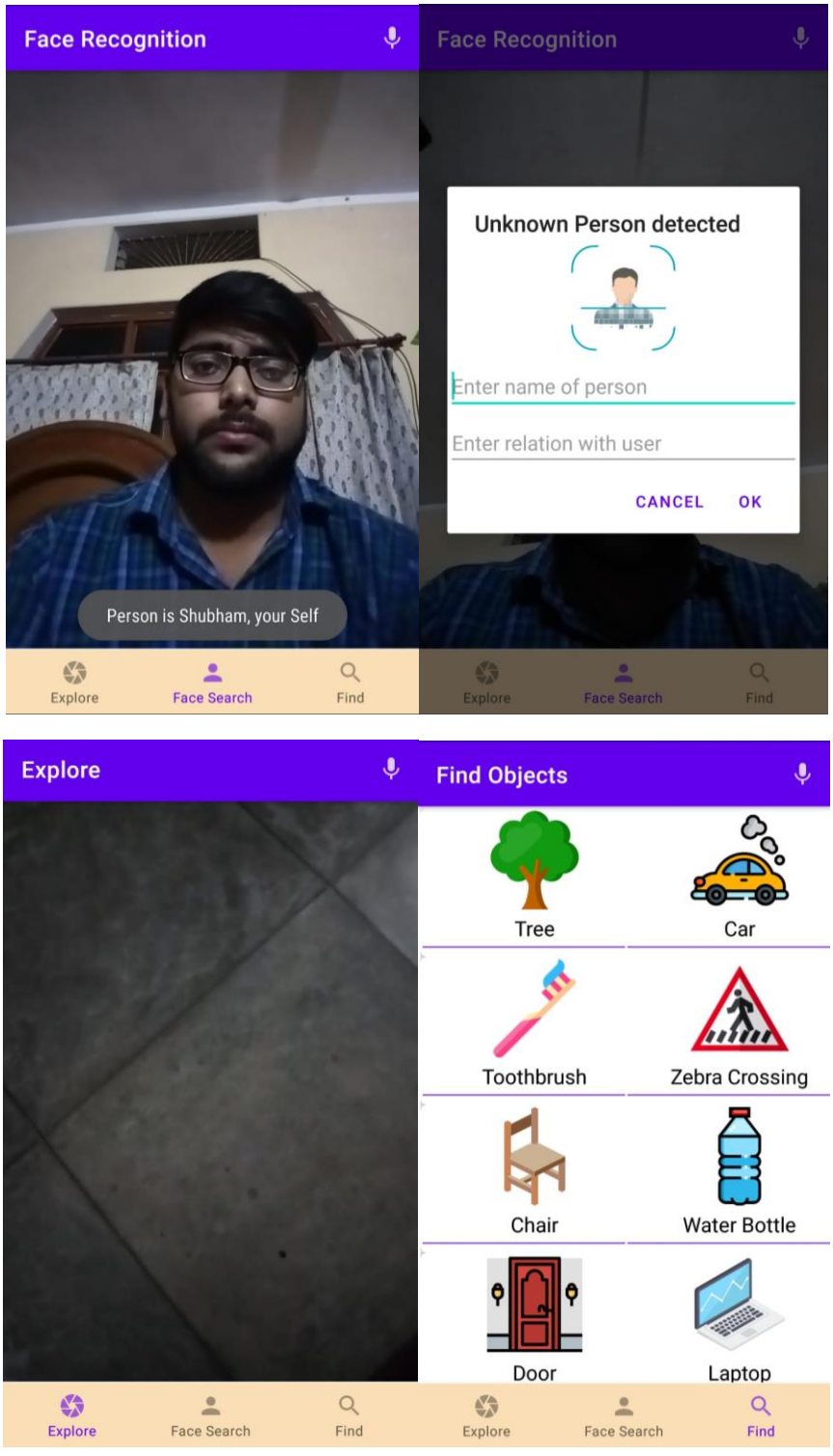

Figure 17.1: System Screenshots

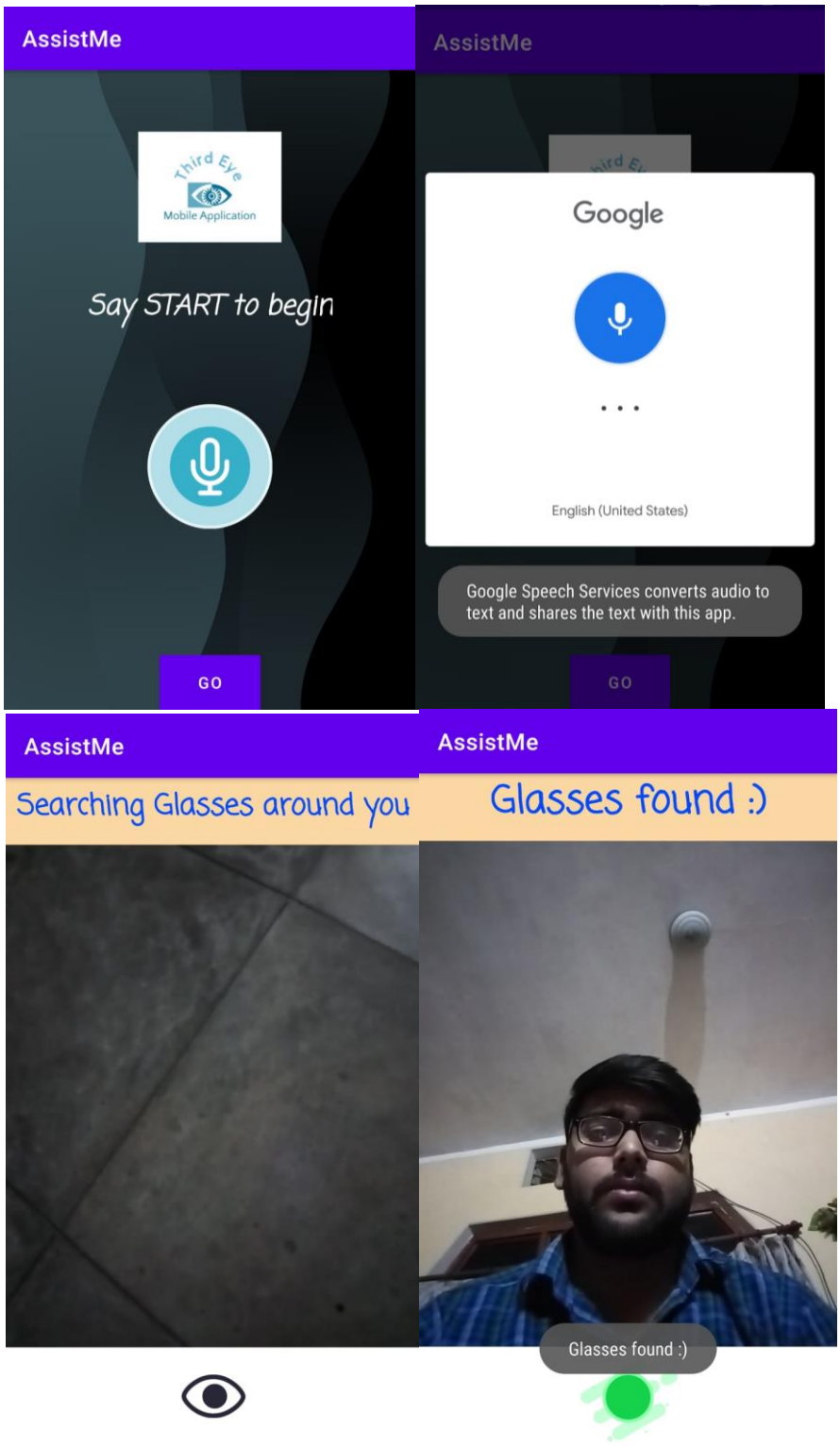

Figure 17.2: System Screenshots

5.4 Testing Process

5.4.1 Test Plan

5.4.1.1 Features to be tested

Ongoing with the workflow, the features to be tested were the capturing of the video using the camera of the computer for initial testing and then proceeding to the video capturing using mobile camera. Next it is to test for the detection of all different kinds of objects. Once the object is detected, we need to test the feature of object recognition and listing and tagging of objects in accordance with the application algorithm. Apart from these, the last feature to be tested is the facial detection and recognition, along with querying the database for person name and relation.

5.4.1.2 Test Strategy and Techniques

White Box Testing:

The list of modules is as follows -

Model 1: Main_Gui: Creates the Graphical User Interface and handles the on-screen events
Input: User generated events like button clicks and voice commands
Output: Calling of other modules depending on the event generated

Model 2: Detect_Object: Processes and trims the video to keep only the motion part of video
Input: Any recorded/captured video in mp4 or mpeg format
Output: Details of detected objects

Model 3: Search_Object: Gives a list of objects detected in video and finds the selected object in a specific video
Input: A video from which objects are to be detected and a csv file having a list of all detected objects in the video and a user generated event to select the object to be searched from a list
Output: Those parts of video pasted together which have the selected object.

Module 4: Capture_video: Starts video capturing and processing with motion detection
Input: A camera interface to record video and start command
Output: Video being recorded and shown on screen simultaneously

Model 5: Send_Voice_Command: It turns the audio to text and then to a particular command which then operated by the application
Input: An audio file that contains the command in mp3 format
Output: Name of the command or the function name

Model 6: Voice_Reply: It turns the text to audio and then it plays it to the user's headphone
Input: A text file
Output: An audio file that contains the input text in mp3 format in English language

In Black Box Testing, we passed the model with different varieties of objects to check if they are detected or not. Objects like cars, buses, bikes, people, birds, cats, cows, dogs, horses, sheep and other animals.

5.4.2 Test Cases

We have taken the following test cases to be tested against the system developed by us :-

1. No objects - This category of video samples contains no objects.

2. Minimal objects - This category of video samples contains just a few objects.

3. Moderate objects - This category of video samples contains moderate quantities of objects.

4. Maximum - This category of video samples contains a large quantity of objects.

5.4.3 Test Results

1. The features being tested are working correctly.

2. The mobile camera is able to capture video streaming. On processing the video for each test case of the four, appropriate results are obtained.

3. The objects are properly detected and communicated to the user through the headphones.

4. The voice commands are working properly.

5.5 Results and Discussions

We have completed designing and programming of all the main modules along with the second phase of machine model training and deployment, all the modules are working fine. We have been able to detect the objects, people, distinguishing them into known and unknown, send audio replies into the user's headphones, and get verbal commands from the user & processing it. Now the user is able to view the scenario around with the help of this virtual eye. The use of this software is not restricted to a specific area but can be used in a variety of different scenarios. The areas where this is useful includes any parking lot of different kinds of vehicle, any specific location where a surveillance camera is installed, a jewellery store and many more. Thus, Third Eye is a full-fledged product, which can be well utilised for a huge domain.

5.6 Inferences Drawn

Object Detection and computer vision is a fast-developing field of work and a lot of research is going on this field. These technologies have widespread applications in various areas. With improvements in the accuracy of the results out of these technologies the scope of these technologies will further expand. By involving technology and making the devices smart we are able to significantly make an impact on costs and time utilised by the product. Initially the development of

the product may take more costs and other investments but it is very much profitable and viable in the long term. This is not just beneficial to the product developer but also to the consumer because it promotes ease of use.

5.7 Validation of Objectives

1. Development to a product which will help users to see the world using their verbal ability.
2. The application is able to detect people, objects near the user and give verbal information of the scenario ahead.
3. The application is able to distinguish between the known and unknown person in front of the user, if the person is known, it spells out his/her name and relation to the user through the headphones/speakers

CONCLUSION AND FUTURE SCOPE

6.1 Work Accomplished

We have completed the first phase of designing and planning along with the second phase of machine model training and deployment, And are now working towards integrating the different models into one user friendly application.

6.2 Conclusions

The overall intention for people with visual impairments is to allow them to experience their surroundings and turn out to be as unbiased as possible. Therefore, we are designing a blind navigation system that is portable and capable to assist them pass the roads and avoid accidents on their outdoor tours. We can say that the system is on its way to turn out to be a practicable answer for assistance and navigation for the blind and partially sighted individuals. The method will increase their self-assurance enabling them to move without help in outdoor spaces and getting to be a more normal person within the society.

We have noticed how important accuracy of results is in this project, hence that will be one of our main focus for this project moving forward.

6.3 Reflections

1. We have gained a wide wealth of information about making the Software Requirement Specification on a realistic project including how to create a plan and a layout for the future tasks to be performed.

2. Through building a prototype we have understood about video codecs and container formats and various technologies like RCNN in object detection and detection with focus on python and usage of Tkinter, along with hardware software integration.

6.4 Future Work Plan

All the basic functionalities and modules are working properly, here are some more steps we can take for the betterment of our application:
1. Code Optimization for improving real time sensing.
2. Improvement of the user interface and user experience of the software.
3. Total hands-free control through voice support.
4. Making it more generic, so that the application can be used for different customers according to their needs.

PROJECT METRICS

7.1 Challenges Faced

1. During the initial phase of the experimentation, the video format obtained using the webcam of the computer system generated only .avi videos. We tried to convert the format of the video to .mp4 so that it can be generalised to a more general format. After some attempts, the final processed video is also stored in the same format that was sent as the input for its processing.

2. Challenges in visual object recognition (at the time) namely: viewpoint variation, illumination and background clutter. In the dataset of the predefined model. It contains objects cleanly centred in the image with minimal background clutter, illuminations, or viewpoint variations. Invariant object detection was yet another issue. Bag of words was used for it. Another dataset that had around 20 types of cars and a few other object classes, pictured from many different angles, scales, and against significant background clutter was used for testing. This dataset helped to systematically study the effects of these various nuisances.

3. Setup of Google Cloud Speech API, we ran into this error after adding the API through the Google console to an existing Google Service Account with JWT credentials. Then we followed the link to the Quickstart Protocol and were able to get it working, then installed a JSON parser to save the JSON file to a secure place and imported this file into the MiaRec application.

7.2 Relevant Subjects

1. Computer Vision

Computer vision is an interdisciplinary study that studies how computers may be programmed to better recognise aberrant states from digitised images or recordings. From the standpoint of construction, it attempts to computerise tasks that the human visual framework can perform.

Computer vision tasks include ways for obtaining, handling, researching, and comprehending computerised images, as well as the extraction of high-dimensional information from this current reality with the objective of generating numerical or representational data, such as in decision-making.

Understanding in this context entails transforming visual images (the retina's contribution) into world depictions that can interact with other points of view and elicit appropriate behaviour. The unravelling of iconic data from visual information using models established with the use of geometry, material science, measurements, and learning assumptions can be considered as picture comprehension.

Computer vision is concerned about the hypothesis underpinning counterfeit frameworks that extract data from images as a logical control. The image data can be organised in a variety of ways, such as video groupings, views from many cameras, or multidimensional data from a medical scanner. Computer vision, as an inventive order, seeks to apply its theories and models to the development of computer vision frameworks.

Scene recreation, event discovery, video following, question recognition, 3D present estimate, grasping, ordering, movement estimation, and picture reclamation are all subspaces of Computer vision.

Computer vision is an interdisciplinary area that deals with how computers may be programmed to better recognise aberrant states from digitised images or recordings. From a design standpoint, it appears to automate tasks that the human visual framework can perform. "Computer vision is concerned with the automated extraction, evaluation, and comprehension of useful data from a single image or a collection of images. It entails refining a hypothetical and algorithmic premise in order to achieve programmed visual comprehension." Computer vision is concerned by the notion underpinning counterfeit frameworks that separate data from pictures as a logical control. The picture data might be in a variety of formats, such as video sequences, images from many cameras, or multidimensional data from a restorative scanner. Computer vision, as a mechanical control, seeks to apply its theories and models to the creation of computer vision frameworks.

2. Image Processing

Image processing is a method of performing a series of operations on a photograph with the purpose of improving the image or removing some valuable data. It's a kind of flag preparation in which the input is a photo and the yield is a picture or its attributes/highlights. Picture processing is one of the most rapidly evolving developments these days. It also has an impact on the centre of research in the fields of design and software engineering.

Picture preparing fundamentally incorporates the accompanying three stages:
i. Bringing in the picture by means of picture securing instruments.
ii. Examining and controlling the picture.
iii. Yield in which the result can be adjusted to a picture or report that depends on picture examination.

There are two types of strategies used for image preparation, simple and computerised image handling. Simple image manipulation can be used for printed versions such as printouts and photos. Picture experts use various essentials of comprehension while utilising these visual systems. Advanced image preparation procedures aid in the control of computerised images through the use of PCs. Pre-handling, upgrading, and data extraction are the three general stages that a wide range of information must go through when using a computerised system.

In this address, we will go over a few key definitions, such as picture, computerised picture, and advanced picture preparation. Different wellsprings of advanced pictures will be discussed, and precedents for each source will be provided. This address will shroud the continuum from image preparation to PC vision. Finally, we will go over image acquisition and various types of image sensors.

7.3 Peer Assessment Matrix

TABLE 5: Peer Assessment Matrix

		Evaluation of		
		Utkarsh	**Shubham**	**Bineet**
Evaluation By	**Utkarsh**	5	5	5
	Shubham	5	5	5
	Bineet	5	5	5

7.4 Role Playing and Work Schedule

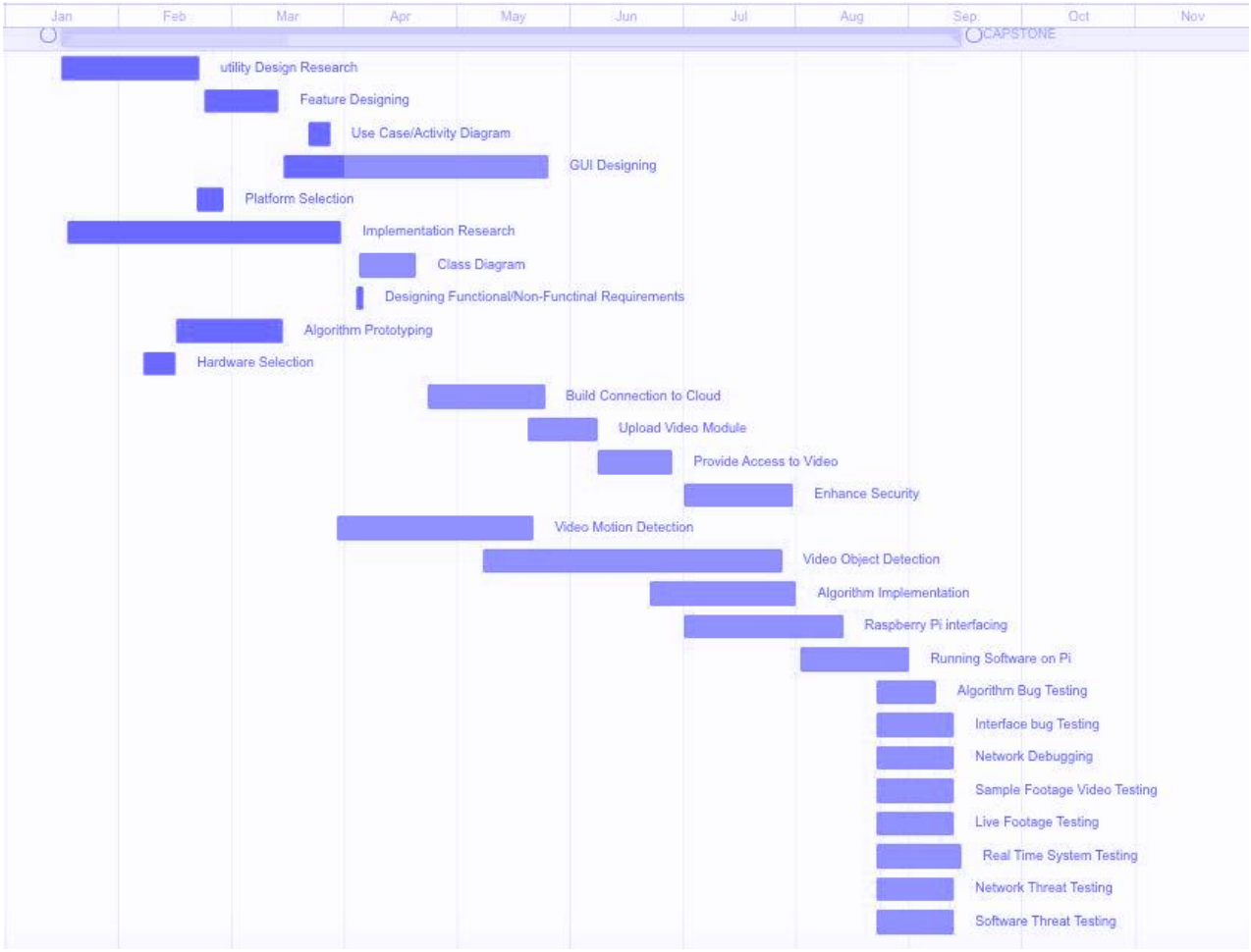

Figure 18: Role Playing and Work Schedule

7.5 Student Outcomes Description and Performance Indicators (A-K Mapping)

The A-K Mapping, includes various parameters that were involved throughout the project and their respective outcomes.

TABLE 6: A-K Mapping

SO	Description	Outcome
1.1	Applying mathematical concepts to obtain analytical and numerical solutions.	An ability to apply knowledge of mathematics, science and engineering.
1.2	Applying basic principles of science towards solving engineering problems.	An ability to apply knowledge of mathematics, science and engineering.
1.3	Applying engineering techniques for solving computing problems.	An ability to use techniques, skills and modern engineering tools necessary for engineering practice.
2.1	Identify the constraints, assumptions and models for the problems.	An ability to design a system, component or a process to meet desired needs within a realistic constraint such as economic, environmental, social, political, ethical, health, safety, manufacturability and sustainability.
2.2	Use appropriate methods, tools and techniques for data collection.	An ability to formulate, identify and solve engineering problems.
2.3	Analyse and interpret results with respect to assumptions, constraints and theory.	An ability to design a system, component or a process to meet desired needs within a realistic constraint such as economic, environmental, social, political, ethical, health, safety, manufacturability and sustainability.
3.1	Design software systems to address desired needs in different problem domains.	An ability to design and conduct experiments, as well as to analyse and interpret data.
3.2	Can understand scope and constraints such as economic, environmental, social, political, ethical, health and safety, manufacturability, and sustainability.	An ability to design a system, component or a process to meet desired needs within a realistic constraint such as economic, environmental, social, political, ethical, health, safety, manufacturability and sustainability.
4.1	Fulfil assigned responsibility in multidisciplinary teams.	An ability to function on multi-disciplinary teams.

4.2	Can play different roles as a team player.	A willingness to assume leadership roles and responsibilities.
5.1	Identify engineering problems.	The broad education is necessary to understand the impact of engineering solutions in a global, economic, environmental and societal context.
5.2	Develop appropriate models to formulate solutions.	An ability to apply knowledge of mathematics, science and engineering.
5.3	Use analytical and computational methods to obtain solutions.	An ability to apply knowledge of mathematics, science and engineering.
6.1	Showcase professional responsibility while interacting with peers and professional communities.	To understand professional and ethical responsibility.
6.2	Able to evaluate the ethical dimensions of a problem.	To understand professional and ethical responsibility.
7.1	Produce a variety of documents such as laboratory or project reports using appropriate formats.	An ability to apply knowledge of mathematics, science and engineering.
7.2	Deliver well-organised and effective oral presentations.	An ability to communicate effectively, both orally and writing.
8.1	Aware of the environmental and societal impact of engineering solutions.	An ability to design a system, component or a process to meet desired needs within a realistic constraint such as economic, environmental, social, political, ethical, health, safety, manufacturability and sustainability.
8.2	Examine economic trade-offs in computing systems.	An ability to use techniques, skills and modern engineering tools necessary for engineering practice.
9.1	Able to explore and utilise resources to enhance self-learning.	A recognition of the need for, and an ability to engage in lifelong learning.
9.2	Recognize the importance of life-long learning.	A recognition of the need for, and an ability to engage in lifelong learning.
10.1	Comprehend the importance of contemporary issues.	A knowledge of contemporary issues.
11.1	Write code in different programming languages.	An ability to design and conduct experiments, as well as analyse and interpret data.

11.2	Apply different data structures and algorithmic techniques.	An ability to communicate effectively, both orally and writing.
11.3	Use software tools necessary for computer engineering domain	An ability to use the techniques, skills and modern engineering tools necessary for engineering practice.

REFERENCES

[1] A. Amer, “A computational framework for simultaneous real-time high level video representing”, Multisensor Surveillance Systems, pp 149-182, 2003.

[2] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, et al. 2012. Video in sentences out. arXiv preprint arXiv:1204.2742,(2012).

[3] A. Kojima, T. Tamura, and K. Fukunaga. 2002. Natural language description of human activities from video images based on concept hierarchy of actions. IJCV 50, 2 (2002), 171-184.

[4] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele. 2017. Movie description. IJCV, Vol. 123, 1, 94-120.

[5] Bhupinder Singh, Neha Kapur, Puneet Kaur “Speech Recognition with Hidden Markov Model: A Review” International Journal of Advanced Research in Computer and Software Engineering, Vol. 2, Issue 3, March 2012.

[6] Jingdong Chen, Member, Yiteng (Arden) Huang, Qi Li, Kuldip K. Paliwal, “Recognition of Noisy Speech using Dynamic Spectral Subband Centroids” IEEE SIGNAL PROCESSING LETTERS, Vol. 11, Number 2, February 2004.

[7] J.W.Davis and A.Tyagi, “A reliable inference framework for recognition of human actions”, Advanced Video and Signal Based Surveillance, 2003.

[8] J.Yamato et al., “Human action recognition using HMM with category separated vector quantization”, IEIC, 1994.

[9] L. M. Fuentes and S. A. Velastin, “People tracking in surveillance applications,” Proceedings 2nd IEEE International Workshop on PETS, Kauai, Hawaii, USA, December 9, 2001.

[10] M.Hamid, “ARGMode-Activity recognition using graphical models”, CVPR, 2003.

[11] R. T. Collins, A. J. Lipton and T. Kanade, “A system for video surveillance and monitoring” Proc. American Nuclear Society (ANS), Eighth International Topical Meeting Robotic and Remote Systems, 1999.

[12] S. Hongeng et al., “Video-based event recognition: activity representation and probabilistic recognition methods”, Computer Vision And Image Understanding Volume: 96 Issue: 2 Pages: 129- 162, 2004.

[13] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. 2015. Sequence to sequence-video to text. In IEEE ICCV.

PLAGIARISM REPORT