CHAPTER 1

INTRODUCTION

1.1 About Deep Learning

Deep learning is a subset of machine learning, which is a subset of artificial intelligence. Deep learning algorithms attempt to draw similar conclusions as humans would by continually analyzing data with a given logical structure. To achieve this, deep learning uses a multi-layered structure of algorithms called neural networks.

Deep learning algorithms attempt to draw similar conclusions as humans would by continually analyzing data with a given logical structure. To achieve this, deep learning uses a multi-layered structure of algorithms called neural networks. The design of the neural network is based on the structure of the human brain. Just as we use our brains to identify patterns and classify different types of information, we can teach neural networks to perform the same tasks on data. The individual layers of neural networks can also be thought of as a sort of filter that works from gross to subtle, which increases the likelihood of detecting and outputting a correct result. The human brain works similarly. Whenever we receive new information, the brain tries to compare it with known objects. The same concept is also used by deep neural networks.

Neural networks enable us to perform many tasks, such as clustering, classification or regression.With neural networks, we can group or sort unlabeled data according to similarities among samples in the data. Or, in the case of classification, we can train the network on a labeled data set in order to classify the samples in the data set into different categories.In general, neural networks can perform the same tasks as classical machine learning algorithms (but classical algorithms cannot perform the same tasks as neural networks). In other words, artificial neural networks have unique capabilities that enable deep learning models to solve tasks that machine learning models can never solve.

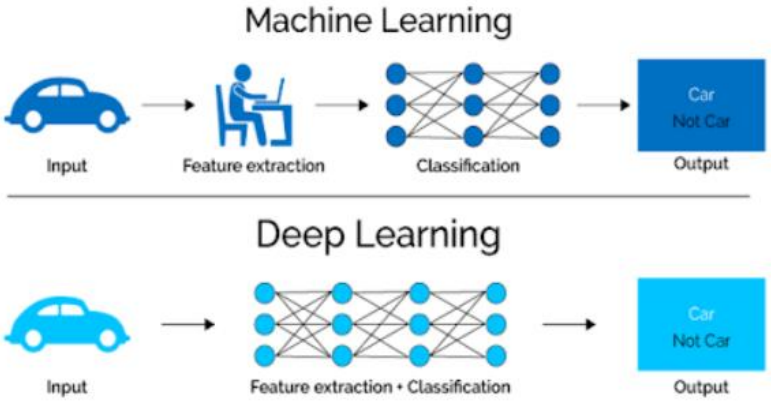

Fig 1.1 Deep Learning

All recent advances in artificial intelligence in recent years are due to deep learning. Without deep learning, we would not have self-driving cars, chatbots or personal assistants like Alexa and Siri. Google Translate would continue to be as primitive as it was 10 years ago before Google switched to neural networks and Netflix would have no idea which movies to suggest. Neural networks are behind all these technologies. A new industrial revolution is taking place, driven by artificial neural networks and deep learning. At the end of the day, deep learning is the best and most obvious approach to real machine intelligence.

1.2 About Object Detection and Text Recognition

Deficiencies in the visual system may contribute to visual impairment which can lead to blindness in the worst cases, which may prohibit individuals from performing many day today tasks, including learning, work, and even walking. According to the World Health Organization around 38 million people worldwide suffer from blindness, while the other 110 million have other types of defects. Recent statistics show that multiple degrees of blindness affect 7 in 1,000 people,with a total world population of 5.3 billion. Unfortunately, in developing countries there are over 90 percent of people suffering from blindness.

Under Vision 2020, India will lessen the predominance of visual deficiency to 0.3% of the all-out populace. India presently has around 12 million visually impaired individualsagainst 39 million all around, which makes India home to 33% of the world visually impaired populace. Iraqis where the visual impedance is a huge issue because of them across the board fear based oppressor action and inherent variations from the

norm in infants brought about by water and nourishment contamination. All things considered, innovative advances havepermitted help to be given to the individuals who live in unfortunate conditions. Subsequently, visually impaired individuals are commonly ready to perform day-to- day tasks independently such as driving through the streets and traveling inside houses.In addition to increased autonomy, often individuals with visual difficulties require helpto identify challenges and are generally supported by other trusted people. However previous research has suggested many strategies to overcome the issues of visually impaired people (VIPs) and offers higher mobility, but these strategies have not been able to fully address the safety measures when VIPs walk on their own. However, the proposed ideas haven't indicated a mechanism for blind people to be in constant interaction with their loved ones and are generally high in complexity and not cost effective. In this paper, a system is proposed that aids visually impaired people in dealingwith day-to-day activities like walking, working, doing house chores and reading texts.This system includes three modules, namely object detection, lane segmentation and finally text recognition. The system comprises of a device by the visually impaired, whereby the device includes a camera module and speaker. The camera will capture theobject's image that is in front of the person, thereafter it gets processed using machinelearning methods and in turn the output which is the name of the object will be converted into audio for the user through the speaker.

Object detection mainly deals with identification of real-world objects such as people, animals, and objects of suspense or threatening objects. Object detection algorithms usea wide range of image processing applications for extracting the object's desired portion. It is commonly used in applications such as image retrieval, security, Medical field, and defense.

1.3 Problem Statement

This project will help a blind person to walk easily by finding the path, detect object in front of them and recognize the objects in the image and output the result using the audio feedback system. It will help them read texts as well and converts text to speech, which is the final output for the users.

1.4 Objectives

- To capture images, and the YOLO (You only look once) algorithm recognize the objects in the image and output the result using the audio feedback system.
- It recognizes the textual data and converts it into speech.

1.5 Scope of the Project

In this computing era, image processing has spread its wings in human life up to the extent that image has become an integral part of their life. There are various applications of image processing in the field of commerce, engineering, graphic design, journalism, architecture and historical research. In this work, Image processing is considered for detecting the object and recognizing the text in front of the visually impaired people.

The study on object detection and recognition can be a helpful aspect in finding the obstacles in front of visually impaired people.

1.6 Motivation

Deficiencies in the visual system may contribute to visual impairment which can lead to blindness in the worst cases, which may prohibit individuals from performing many day today tasks, including learning, work, and even walking. Computer vision is one such field which has millions of possibilities and this project itself being a primary example of it. This project aims to help the blind society to experience the world independently with the help of a speech-based feedback device.

1.7 Organization of the Report

In chapter 1 we have discussed about the introduction to Machine learning, Problem statement and objectives. In the chapter 2 we discuss about the literature survey consisting of various related technical papers. In chapter 3 we discuss about the System Analysis. In chapter 4 we discuss about the System requirements and specification. In chapter 5 we discuss about the system architecture. In chapter 6 we discuss about the System implementation, modules, and pseudo codes. In chapter 7 we can see the snapshots and results of the project. At last we discuss about the conclusion and future work.

CHAPTER 2

LITERATURE SURVEY

Literature survey review refers to the content getting from the books which is related to the topic. It should be referred from some research paper which is related to the topic which is given to the student. Any material which is related to the paper from internet and which is valuable for student and that literature review helped the student to enhance the report status and calculation, analysis and tabulation also strong which majorly reflects in the report.

2.1 Related Work

The paper [1], presents a synopsis of enabling a real-world experience through a wearable speech-based feedback system. The idea of a wearable device that includes a Raspberry Pi 4 and camera module to provide feedback to signal obstacles to the users and also identify and read out texts is proposed. When a text is placed in front of the camera module, the text is first recognized and then read out to the user. Similarly, objects present in front of the user are identified and communicated to the person who is wearing the device. In a study conducted, it was found that visually impaired people had difficulty in identifying whether there are any hindrances in front of them or what textual content is present in front of them and hence this project. This project solves these challenges and aids visually impaired people to get their tasks done in the same manner as that of a normal person. Therefore is aiming to make the living of visually impaired people easier as well as help them get through their daily activities without coming in contact with any dangerous obstacles and wish to incorporate several new features to the system like navigation, which helps the user with directions, like going left or right, and also wish to include object detection in dim light, which is a crucial feature as without it visually impaired people would be living a restricted life.

The author of the paper [2], aims at building a software that can detect and classify and extract color of the object towards which the camera is pointed. This research accomplishes this by using state of the art object detection algorithm called YOLO (You only look once). The color extraction is performed by detecting the different hues in the image, clustering them accordingly using K- means and then considering the color that has thelargest cluster size with the background of the image masked. The software's

camera module takes in the object and after comparison with the dataset; it recognizes the objectin the image frame if any. After recognition, the audio synthesizer helps use give an audio output as audio cues are the only way the detection can be conveyed to the wearer.The software will contain the ability to detect suspicious objects like knives, guns, etc. A warning will be given to the wearer about the suspicious object via audio feedback as well. With this paper being able to detect suspicious objects it'll be of help to the wearerof the device who'll be a person with impaired vision to be more aware of his/her surroundings making him feel secure by our research. This paper plays a very importantrole in making the visually impaired people feel in a safe environment. The software for color blind people provides with color information as well proving it to be of immense help to color blind users of the software. The delay between the detection of the object and the audio feedback to the user is made the minimum and the audio feedback is made clear enough for the user to be understood.

The paper[3], has proposed an IoT-enabled automated system that can help visually impaired in their safe navigation and identifies several common objects in indoor and outdoor environments in real-time scenarios. The currency notes, currently used in Bangladesh, can also be recognized by the proposed system. The objects are detected by laser sensors in the direction of the front, left, right, and ground. The SingleShot-Detector is used to recognize objects and an audio signal is provided by thesystem to aware the user using the headphone through Bluetooth technology. The pre-trained model is prepared on a host computer using two custom datasets that are created manually. Overall, the system can detect all types and shapes of objects but can recognize five different types of objects and eight currency notes. The object detection and recognition accuracy of the developed system is 99.31% and 98.43% respectively. The developed system can send a warning notification to others (friends, relatives, etc.) in case a true free fall has been occurred. All processed data and information regarding objects' recognition are sent and stored to a remote server in real-time to ensure the safety of the users and for further analysis by the researchers. At present, the weight of the system is a few hundred grams and all the components are connected by wires that have increased the size of the system. Also, the system can recognize five different types of objects with eight currency notes. The future enhancements of the proposed method may focus on the development of a system-on-chip (SoC) so that thesize, weight, and cost of the system can be reduced. Moreover, new methodology could

be introduced to increase the number of objects to be recognized.

The author of the paper [4], present a visual substitution system for blind people based onobject recognition in video scene. This system uses SIFTS key points extraction and features matching for object identification. They devote the experimental part to test the application in order to detect some objects in some video scene with different conditions. In this stage of works, we address the recognition of each object in the scene as an individual task; they do not consider the relationships between many objects. Thus, in future works, we will consider this relationship for scene understanding or detecting everything that belongs to a given place or location. Finally, in order to help bind people and to provide from the new technologies, a mobile application can be the best solution.

The paper [5], propose a wearable smart electronic system used as aid for VIPs. The novelty of the system is the remote monitoring capabilities available to family and caregivers, the detection of the user stumbling over, and an alarm to request assistance. Moreover, we integrated these functionalities into a wireless wearable system endowed with sensors, alarms, power supply, cellular module, and GPS. The implemented system is intended for users of any age. The system can be used by VIP people, even by those with hearing impairments. In addition, the system architecture provides improved efficiency compared to similar systems. For instance, the solar panel provides additional support for power supply and it is useful to charge batteries when navigating outdoors. Likewise, the system is endowed with different types of alarms that send warnings to the user, surrounding people, family, and caregivers.

The paper [6], proposed several technologies used for detecting obstacles for blind persons have been reviewed. The advantages and disadvantages of previous work have been studied thorough along with their critical challenges and future directions. The analysis and comparison of several techniques have been done, which may serve as research Gaps for upcoming researches in this field. After reviewing many papers on this work, it is observed that major improvements are required in existing systems, so that they can operate accurately in crowded areas such as supermarkets, hospitals, airports etc. Also, some techniques yield accurate results but takes more timeduring execution, thereby making the systems unfit for real time situations. Different combination and placements of sensors along with user's learning curve will provide research platform in making different assistive tools for Visually Impaired (VI) users. Motivating blind users towards new technology itself impose a big challenge. This

paper will help in the ongoing research on this topic.

The author in the paper [7], presents the miniaturization of actuators and electronics has allowed the creation of new devices and systems that can be embedded into clothing. These wearablesystems facilitate the user's ability to perform normal daily tasks without feeling encumbered by burdensome devices. In particular, this chapter has focused on wearableassistive devices for the blind. A brief non-exhaustive survey of wearable assistive devices for this population has been presented to illustrate the most representative workdone in this area. Devices worn on the finger, hands, wrist, fore arm, tongue, head, chest,abdomen and feet have been proposed over the last decades to provide wearable solutions to the problems of reading and mobility. For the blind, hearing and touch become the first and second major senses, respectively. They will never replace visionbut they still gather much information from the environment for daily tasks. That is thereason why assistive devices provide acoustical and tactile feedback to compensate forvisual information. In contrast, smell and taste are largely ignored as being essential to the interaction with the environment. Several universal design concepts for acoustical/tactile based assistive devices have been presented. They provide guidelines to stimulate both hearing and touch in order to obtain the best performance from these senses.

The author in the paper [8], proposed the various object detection algorithms such as skin detection, color detection, face detection and target detection are simulated using MATLAB 2017b with an accuracy of approximately 95%. Parameters such as detection accuracy, RGB Euclidian Threshold ‘T’ in Target Detection, Y, Cb and Cr in Skin Detection have been simulated and implemented to improve the efficiency of the algorithms for video surveillance applications. Further a single algorithm maybe designed by considering various detection parameters such as Color, Face, Skin and Target of interest to meet video surveillance applications.

CHAPTER 3

SYSTEM ANALYSIS

3.1 Existing System

The computer vision system now a day is normally consist of computer, digital camera and application software. Various types of algorithms are integrated in the application. Image processing is one important method that helps segments image into objective and background image.

Various approaches have been proposed in the literature for the design of visual assistance systems based on the underlying sensory systems, hardware configuration, physical setup, data inference techniques and user interface. The most used sensor types include ultrasound, sonar, laser, RGB CCD camera, infrared (IR) camera and GPS. Some approaches convert the input sensor data to other modalities. For the user interface, audio transmitted via earphones or hand gloves equipped with buzzers or tiny vibrating actuators are typically used.

Early visual assistance system designs were based on projecting a camera image onto the human skin using vibrating motors and sensor modality conversion, where ultrasonicwaves were converted to the audible range and the converted audio was used tounderstand the environment. Although the voice system, where visual image data are converted to human audible frequency, showed promising results these systems were typically slow, physically uncomfortable, and obtrusive, provided only very coarse information about the surroundings and required extensive user training to be used effectively. Early visual assistance systems based on GPS data, focused primarily on navigation (i.e., neither collision avoidance nor obstacle detection was performed) and often suffered from signal loss, especially in indoor environments and urban areas. Visual assistance systems based on RFID technology provided good localization in indoor environments where in RFID tags were physically placed. Since RFID sensing methods provide a range rather than an accurate geolocation of the tags, the resulting localization errors were unacceptable in certain situations.

3.2 Disadvantages of Existing System

- Most of the aids for VIPs rely on sensors for obstacle detection using different

sensing principles and devices. For instance, the authors of used different sensors to detect obstacles, namely, sonar, infrared, and ultrasonic sensors, and the same localization (i.e., GPS) and communication (i.e., GSM) technologies. GSM is considered as an affordable solution, but its performance decreases indoors.

- Nearly all existing systems are focused in limited directions.
- Existing systems may contain errors.
- Low accuracy.

3.3 Proposed System

The proposed system includes a device that helps visually impaired people to move around and get their day today tasks done independently like every other person. The device that contains a Web camera module and speaker. It consists of object detection, text recognition and text to- speech conversion. Basic Operations:

- Object Detection
- Text Recognition
- Text-to-Speech

1. Object Detection: Object Detection is a computer technology related to computer vision and image processing that is used in digital images and videos to recognize instances of symbolic objects of a certain type (such as humans, homes or cars). So basically, this is incorporated in our project so that it can help the blind people identify commonly used objects in daily life as well as in identifying walkable space. Whenever there is an object or obstacle in front of the person, it alerts the person and helps them to avoid it. Objects are classified to help the person know of what objects are in front of them using image processing, therefore helping them to get a sense of their surroundings.

In this system, did with the help of OpenCV and the pretrained deep learning model, ie,YOLOv3 to identify objects. The model "YOLO" algorithm is run through various highcomplexity convolutional neural network architectures which is known as the Darknet.Common Objects in Context (COCO) dataset is used to train the model. COCO is pre- trained dataset, so there is no need for external training. OpenCV, keras, and image modules are imported in our python program. The python cv2 package has a

technique to set up Darknet from the configurations in the yolov3.cfg file and capture the live video from the Pi camera and then give it as input to a pre-trained YOLO model whichhas more than 80 objects classified.

The prediction of the class of the objects identified in each frame is a string e.g. “dog”.This will also retain the coordinates of the objects in the image and include the position“top left” etc to the class “dog”. Then the text description is sent to the gTTS API usingthe gTTS package.

2.Text Recognition: It recognizes the textual data placed in front of the camera module with the help of text recognition and converts it into speech, thus helping visually impaired to identify the textual information in front of them.

OCR (Optical Character Recognition) Text recognition with Python and API is used inthis system. OCR is the recognition of a text from an image with the use of a computer.Ocr.spaceis an OCR engine which offers free API. In short OCR does all the work of text detection by sending their API with the selected image with the required text that want to scan and the returned result is the text scanned. step you have to import the required libraries (OpenCV, IO, numpy, requests and json), whereby IO and Json are already installed on python which makes it default, and then load the image. Thereaftercan cut the image with the required area of text incase the image contains some sort ofbackground disturbance. Then the image is to be sent to the ocr. space server to be furtherprocessed and need to compress the image into JPG format so that the image will be sent with a maximum size of 1MB using the free service, and this will shrink the size of the image. Later on, convert the image into bytes, and then send the bytes to the serverusing the python library requests, by sending three parameters. First one is the url-api,second is the called “files” which has the name of the file and the bytes of the file that were generated after the compression of the image. And thirdly is the “Data” which includes the post parameters of the OCR engine. The API key then needs to be insertedwhere it is now written “YOURAPIKEYHERE”, where the language is the language of our text, which is English. Thereafter the function will send the image to the server and will get a response from the server. Finally, the result from the server is a string which is converted into a dictionary.

3.Text-to-Speech: gTTS is used to convert textual contents to audio feedback, which is the final output to the visually impaired user. This system uses gTTS (Google Text-

To-Speech) which is a python module and a CL method to interact with the Google Translate Text-to-Speech API.GTTS is a very easy-to-use device that converts the text that the machine has recognized and then translates the text to audio that is saved in an mp3 format. The gTTS API supports many languages such as Hindi, English, Malayalam, French and many more. It is not possible to change the voice of the audio, however it is possible to change the audio speeds to fast or slow. The first step of the code is to import the gTTS library and the "os" module which is required to hear the converted audio. Thereafter the text is sent to an object called "text", and then an object called "speech" is created to pass the text and required language to the engine. "fast=false" indicates that the text should have slow speed when read out. The object "speech" that contains the converted file will be saved in a mp3 file named 'text.mp3'. To play the converted file, the Windows command "start" is used following the mp3 file name.

3.4 Advantages of Proposed system

- In contrast to present system which emphasize more on user easiness of naive users, this system focus more on user easiness of visually disabled people.
- All operations will be based on detection of the camera module.
- Better equipped to handle dynamic systems.
- Techniques like Background extraction and feature selection are used.

CHAPTER 4

SYSTEM REQUIREMENTS AND SPECIFICATION

4.1 Introduction

A System Requirement Specification (SRS) is a comprehensive description of the intended purpose and environment for software under development. The SRS fully describes what the software will do and how it will be expected to perform.

An SRS minimizes the time and effort required by developers to achieve desired goalsand also minimizes the development cost. A good SRS defines how an application willinteract with the system hardware, other programs and human users in a wide variety ofreal-world situations. Parameters such as operating speed, response recovery from adverse events are evaluated.

4.2 System Requirements

Hardware Requirements

- A system having an i3 or i5 processor and
- A minimum of 8 GB of RAM is ideal for proper functioning.
- Hard disk :1TB
- Windows 7 or above

Software Requirements

- Python-IDLE: Integrated development and Learning environment used in computer programming.

Python -IDLE

Python IDLE (Integrated Development and Learning Environment) is an integrated development environment (IDE) for Python. The Python installer for Windows contains the IDLE module by default. IDLE is not available by default in Python distributions for Linux. It needs to be installed using the respective package managers. IDLE can be used to execute single statement just like Python Shell and also create, modify and execute Python scripts.

IDLE provides a fully-featured text editor to create Python script that includes features like syntax highlighting, autocompletion, and smart indent. It also has a debugger with stepping and breakpoints features.

Advantages of Python- IDLE

- Easy to use and learn
- Python IDLE interface is user friendly
- Syntax highlighting is nice features
- Smart indent helps a lot

Disadvantages of Python-IDLE

- Debugging could be more advanced
- Can have more data science packages
- Output features can be better

4.3 Functional Requirements

In object detection computational vision methodology for the detection of object used by visually impaired people in front of them. Image processing it determines which objects and text in the video scene and recognize the object and text. Similar to the majority of computational vision systems in the calibration stage the system will be responsible for capturing the objects. The objects are subsequently extracted from the video frame the next extracting features selection by using input parameters of computer visionsystem.

CHAPTER 5

SYSTEM DESIGN

Systems Design is a broad term for describing methodologies for developing high quality information system which combines Information Technology, people and Data to support business requirement. Systems Analysis It is a process of collecting and interpreting facts, identifying the problems, and decomposition of a system into its components. System design is the process of defining the components, modules, interfaces, and data for a system to satisfy specified requirements.

5.1 System Architecture

Text Recognition

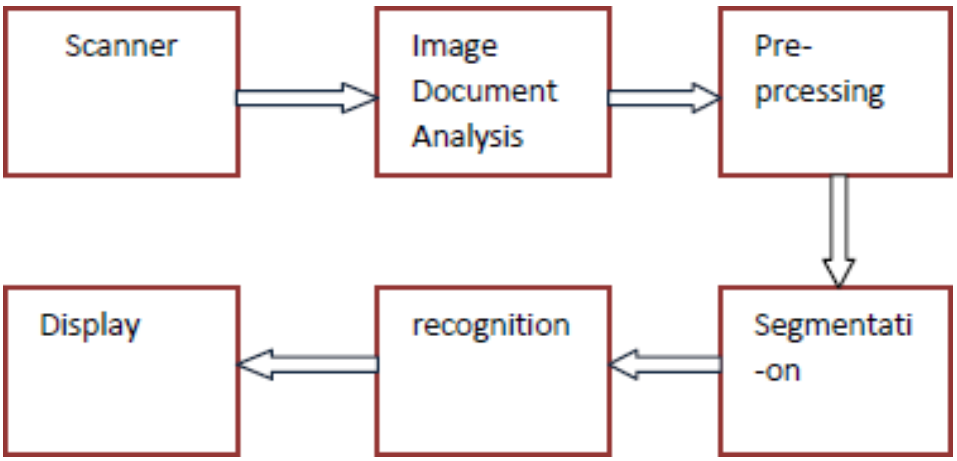

Fig 5.1 Text Recognition

Scanner: An image scanner—often abbreviated to just scanner—is a device that optically scans images, printed text, handwriting or an object and converts it to a digital image. Commonly used in offices are variations of the desktop flatbed scanner where the document is placed on a glass window for scanning. Hand-held scanners, where the device is moved by hand, have evolved from text scanning "wands" to 3D scanners used for industrial design, reverse engineering, test and measurement, orthotics, gaming and other applications. Mechanically driven scanners that move the document are typically used for large-format documents, where a flatbed design would be impractical. Modern scanners typically use a charge-coupled device (CCD) or a contact image sensor (CIS) as the image sensor, whereas drum scanners, developed earlier and still used for the highest possible image quality, use a photomultiplier tube (PMT) as the image sensor.

A rotary scanner, used for high-speed document scanning, is a type of drum scanner that uses a CCD array instead of a photomultiplier. Non-contact planetary scanners essentially photograph delicate books and documents. All these scanners produce two-dimensional images of subjects that are usually flat, but sometimes solid; 3D scanners produce information on the three-dimensional structure of solid objects.

Image Document Analysis: Document image analysis refers to algorithms and techniques that are applied to images of documents to obtain a computer-readable description from pixel data. A well-known document image analysis product is the Optical Character Recognition (OCR)software that recognizes characters in a scanned document. OCR makes it possible for the user to edit or search the document's contents. In this paper we briefly describe various components of a document analysis system. Many of these basic building blocks are found in most document analysis systems, irrespective of the particular domain or language to which they are applied. We hope that this paper will help the reader by providing the background necessary to understand the detailed descriptions of specific techniques presented in other papers in this issue.

Pre-processing: Preprocessing can refer to manipulation or dropping of data before it is used in order to ensure or enhance performance, and is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of- range values (e.g., Income: −100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), and missing values, etc.

Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running any analysis. Often, data preprocessing is the most important phase of amachine learning project, especially in computational biology. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time.

Segmentation: Image segmentation is the process of partitioning a digital image into multiple image segments, also known as image regions or image objects (sets of pixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation

is typically used to locate objects and boundaries (lines, curves, etc.) in images.

More precisely, image segmentation is the process of assigning a label to every pixel inan image such that pixels with the same label share certain characteristics.

The result of image segmentation is a set of segments that collectively cover the entireimage, or a set of contours extracted from the image (see edge detection). Each of the pixels in a region are similar with respect to some characteristic or computed property, such as color, intensity, or texture. Adjacent regions are significantly different color respect to the same characteristic(s). When applied to a stack of images,typical in medical imaging, the resulting contours after image segmentation can be usedto create 3D reconstructions with the help of interpolation algorithms like marching cubes.

Recognition: Recognition is the automated recognition of patterns and regularities in data. It has applications in statistical data analysis, signal processing, image analysis, information retrieval, bioinformatics, data compression, computer graphics and machine learning. Pattern recognition has its origins in statistics and engineering; some modern approaches to pattern recognition include the use of machine learning, due to the increased availability of big data and a new abundance of processing power. These activities can be viewed as two facets of the same field of application, and they have undergone substantial development over the past few decades.

Display: Display is a computer output surface and projecting mechanism that shows text and often graphic images to the computer user, using a cathode ray tube (CRT), liquid crystal display (LCD), light-emitting diode, gas plasma, or other image projection technology. The display is usually considered to include the screen or projection surface and the device that produces the information on the screen. In somecomputers, the display is packaged in a separate unit called a monitor. In other computers, the display is integrated into a unit with the processor and other parts of thecomputer. (Some sources make the distinction that the monitor includes other signal- handling devices that feed and control the display or projection device. However, this distinction disappears when all these parts become integrated into a total unit, as in the case of notebook computers.) Displays (and monitors) are also sometimes called videodisplay terminals (VDTs). The terms display and monitor are often used interchangeably.

Object Detection

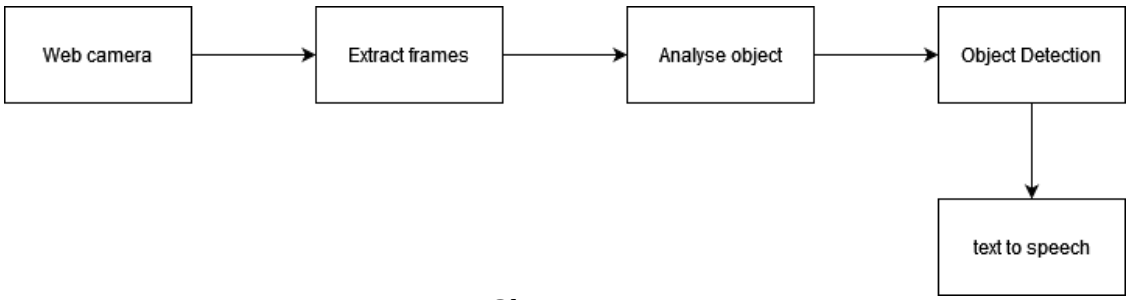

Fig 5.2 Object Detection

Web camera: A webcam is a video camera that feeds or streams an image or video in real time to or through a computer network, such as the Internet. Webcams are typically small cameras that sit on a desk, attach to a user's monitor, or are built into the hardware. Webcams can be used during a video chat session involving two or more people, with conversations that include live audio and video.

Webcam software enables users to record a video or stream the video on the Internet. As video streaming over the Internet requires much bandwidth, such streams usually use compressed formats. The maximum resolution of a webcam is also lower than most handheld video cameras, as higher resolutions would be reduced during transmission. The lower resolution enables webcams to be relatively inexpensive compared to most video cameras, but the effect is adequate for video chat sessions.

Extract frames: Finally, the video frame extraction algorithm is used for the motion-compensated scan conversion of interlaced video data, with a visual comparison to the resolution enhancement obtained from progressively scanned frames.

Import the video which is to be converted into frames into the current matlab environment. Extract the total number of frames in the video. Make an empty directory named frames before the execution. Run a for loop and start extracting the frames into the directory.

Analyse Object: The examination of a problem by modeling it as a group of interacting objects. An object is defined by its class, data elements and behavior. For example; in an order processing system, an invoice is a class, and printing, viewing and totaling are examples of its behavior.

Object Detection: Object detection is a computer technology related to computer vision and image processing that deals with detecting instances of semantic

objects of a certain class (such as humans, buildings, or cars) in digital images and videos. Well-researched domains of object detection include face detection and pedestrian detection. Object detection has applications in many areas of computer vision, including image retrieval and video surveillance.

Text to Speech: Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech computer or speech synthesizer, and can be implemented in software or hardware products. A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech. The reverse process is speech recognition.

Synthesized speech can be created by concatenating pieces of recorded speech that arestored in a database. Systems differ in the size of the stored speech units;a system that stores phones or diphones provides the largest output range, but may lack clarity. For specific usage domains, the storage of entire words or sentences allowsfor high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output.

5.2 Flow Chart

- **Text Recognition**

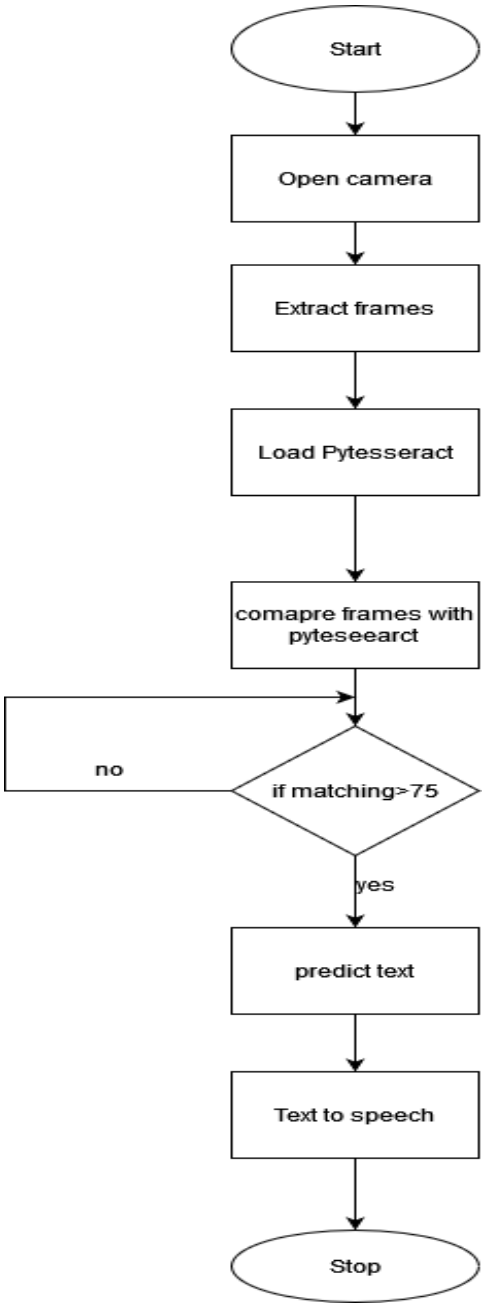

Fig 5.3 Text Recognition Flow Chart

Steps

1. Start the camera
2. Extract the frames
3. Load Pytesseract
4. Compare frames with Pytesseract
5. If comparison is greater than 75% the text will be predicted.
6. Predicted text is converted into speech.

- **Object Detection**

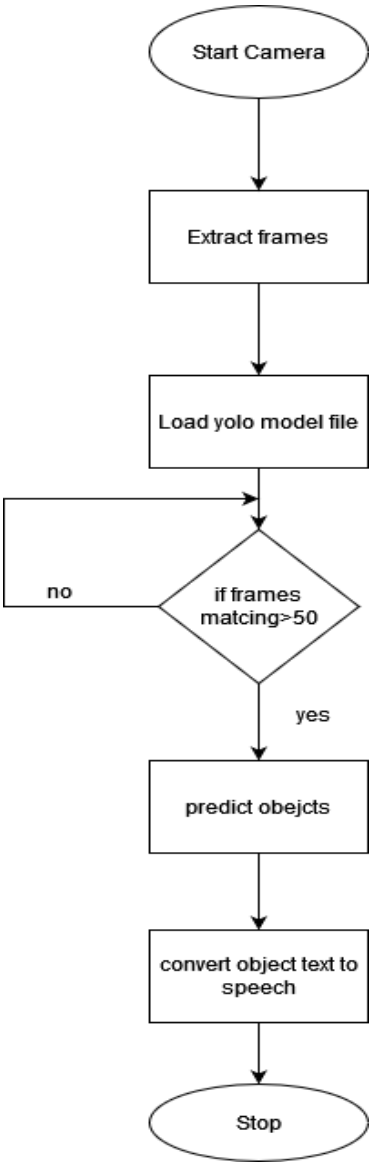

Fig 5.4 Object Detection Flow Chart

Steps

1. Start the camera
2. Extract the frames
3. Load yolo model file
4. Compare frames with yolo model file
5. If comparison is greater than 50% the object will be predicted.
6. Predicted object is converted into speech.

CHAPTER 6

SYSTEM IMPLEMENTATION

Implementation is a process of ensuring that the information system is operational. It involves constructing a new system from scratch. Also, construction of new system from existing one. Implementation is the process of executing a plan or policy so that a concept becomes a reality. To implement a plan properly, managers should communicate clear goals and expectations, and supply employees with the resources needed to help the company achieve its goals.

6.1 Tools Used

The major tool used to accomplish this project was Python IDLE (Integrated Development and Learning Environment) is an integrated development environment (IDE) for Python. The Python installer for Windows contains the IDLE module by default. IDLE is not available by default in Python distributions for Linux. It needs to be installed using the respective package managers.

IDLE can be used to execute a single statement just like Python Shell and also to create,modify, and execute Python scripts. IDLE provides a fully-featured text editor to createPython script that includes features like syntax highlighting, autocompletion, and smartindent. It also has a debugger with stepping and breakpoints features.

6.2 Modules

6.2.1 Text Recognition

Video Input

import os

from pathlib import Path import

sysfrom datetime import datetime

import time import threading

from threading import Thread import cv2

import numpy import pytesseractimport pyttsx3

engine = pyttsx3.init()

```
def tesseract location(root):

try:

pytesseract. pytesseract. tesseract = rootexcept FileNotFoundError:

print ("Please double check the Tesseract file directory or ensure it's installed.")
sys.exit(1)
```

Text Recognition

```
class OCR:

def___init___(self): self.boxes = None self.stopped = False self.exchange = None
self.language = None self.width = None self.height = None self.crop_width = None

self.crop_height = Nonedef start(self):

Thread(target=self.ocr, args=()).start()return self

def set_exchange(self, video_stream):

self.exchange = video_stream def set_language(self, language):

self.language = languagedef ocr(self):

while not self.stopped:

if self.exchange is not None:

frame = self.exchange.frame

frame = cv2.cvtColor(frame, cv2.COLOR_RGB2GRAY)
frame = frame[self.crop_height:(self.height - self.crop_height),

self.crop_width:(self.width - self.crop_width)]

self.boxes=pytesseract.image_to_data(frame,lang=self.language

def set_dimensions(self, width,height, crop_width, crop_height):

self. width = width self.height = height

self.crop_width = crop_widthself.crop_height = crop_heigh

def stop_process(self):

self.stopped = True
```

6.2.2 Object Detection

Video Input

```
import numpy as np

import time

import cv2

import pyttsx3

engine = pyttsx3.init()

LABELS = open("coco.names").read().strip().split("\n")print("[INFO] loading YOLO from disk...")

net=cv2.dnn.readNetFromDarknet("yolov3.cfg","yolov3.weights")

ln=net.getLayerNames()

ln = ln[i - 1] for i in net.getUnconnectedOutLayers()] np.random.seed(42)

COLORS = np.random.randint(0, 255, size=(len(LABELS), 3),dtype="uint8")

cap = cv2.VideoCapture(0)frame_count = 0

start = time.time()first = True frames = []
```

Object Detection while True: frame_count +=

```
1ret, frame = cap.read()

frame = cv2.flip(frame, 1)frames.append(frame)

if ret:

key = cv2.waitKey(1)

if frame_count % 60 == 0:end = time.time()

(H, W) = frame.shape[:2]

blob = cv2.dnn.blobFromImage(frame, 1 / 255.0, (416, 416),

swapRB=True, crop=False)

net.setInput(blob)

layerOutputs = net.forward(ln)boxes = []

confidences = []classIDs = [] centers = []
```

6.3 About COCO Dataset

A good dataset will result in better precision and recall. Some commonly used datasets are ImageNet, Pascal, SUN, and COCO. Yolo is a pre-trained object detection algorithm. The dataset used as training data for the YOLO object detection algorithm is COCO (Common objects in context). COCO is a large-scale detection, segmentation, and captioning dataset. The coco dataset contains around 80 classes, 80,000 training images, and 40,000 validation images.

COCO was an initiative to collect natural images, the images that reflect the everyday scene and provides contextual information. In the everyday scene, multiple objects canbe found in the same image and each should be labeled as a different object and segmented properly. COCO dataset provides the labeling and segmentation of the objects in the images. A machine learning practitioner can take advantage of the labeled and segmented images to create a better performing object detection model.

Fig 6.1 COCO Dataset

6.4 YOLO Algorithm

YOLO is a novel approach to detect multiple objects present in an image in real-time while drawing bounding boxes around them. It passes the image through the CNN algorithm only once to get the output, thus the name. Although comparatively similar to R-CNN, YOLO practically runs a lot faster than Faster R-CNN because of its simpler architecture. Unlike Faster R-CNN, YOLO can classify and perform bounding box

regression at the same time. With YOLO, the class label containing objects, their location can be predicted in one glance. Entirely deviating from the typical CNN pipeline, YOLO treats object detection as a regression problem by spatially separating bounding boxes and their related class probabilities, which are predicted using a single neural network.

This process of performing both bounding box prediction and class probability calculations is a unified network architecture that YOLO initially introduced.

YOLO algorithm extends Google Net equations to be used as their base forwarding transport computation, assumably the reason behind the speed and accuracy of YOLO's real-time object detection. In comparison with R-CNN architectures, unlike running a classifier on a potential bounding box, then reevaluating probability scores, YOLO predicts bounding boxes and class probability for those bounding boxes simultaneously. This optimizes the YOLO algorithm and is one of the significant reasons why YOLO is so fast and less likely to have errors to be utilizable for real-time object predictions.

YOLO's architecture is similar to a typical convolutional neural network inspired by the Google Net model for image classification. The network's initial layer first extracts the image's features, and the fully connected layers predict the output probabilities andcoordinates. With 24 convolutional layers, two fully connected layers, 1x1 reduction layers and 3x3 convolutional layers, the full YOLO network model created.

YOLO modules perform following functionalities:

Frame Analysis: This module allows the model to analyze the capture frame.

Object Registration: This module allows the object to be registered with the model.

Object Detection: This module allows the product to detect the object based on the training received.

Generate Prediction: This module allows the module to generate a text prediction of what the object is.

Text to speech: The module performs the task of conversion of the predicted object to audible form.

6.4.1 How YOLO Algorithm Works

YOLO algorithm works using the following three techniques:

- Residual blocks
- Bounding box regression
- Intersection Over Union (IOU)

Residual Blocks

First, the image is divided into various grids. Each grid has a dimension of S x S. The following image shows how an input image is divided into grids.

Fig 6.2 Residual Blocks

In the image above, there are many grid cells of equal dimension. Every grid cell will detect objects that appear within them. For example, if an object center appears within a certain grid cell, then this cell will be responsible for detecting it.

Bounding box regression

A bounding box is an outline that highlights an object in an image.

Every bounding box in the image consists of the following attributes:

- Width (bw)
- Height (bh)
- Class (for example, person, car, traffic light, etc.)- This is represented by the letter c.
- Bounding box center (bx,by)

The following image shows an example of a bounding box. The bounding box has been represented by a yellow outline.

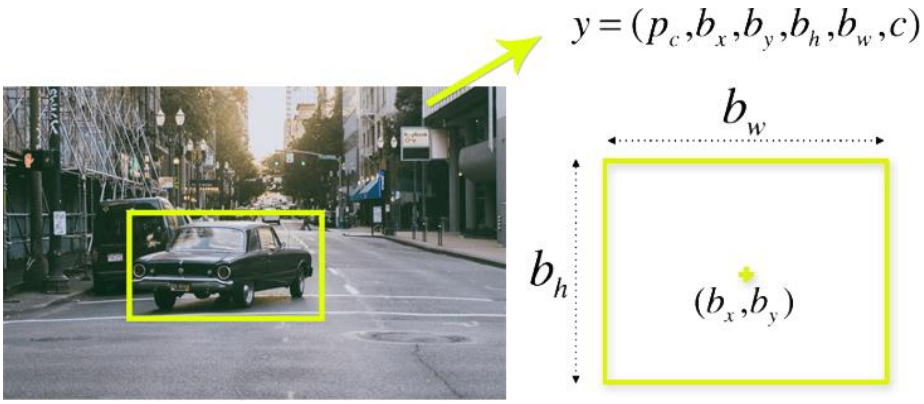

Fig 6.3 Bounding Box Regression

YOLO uses a single bounding box regression to predict the height, width, center, and class of objects. In the image above, represents the probability of an object appearing in the bounding box.

Intersection over union (IOU)

Intersection over union (IOU) is a phenomenon in object detection that describes how boxes overlap. YOLO uses IOU to provide an output box that surrounds the objects perfectly.

Each grid cell is responsible for predicting the bounding boxes and their confidence scores. The IOU is equal to 1 if the predicted bounding box is the same as the real box. This mechanism eliminates bounding boxes that are not equal to the real box.

The following image provides a simple example of how IOU works.

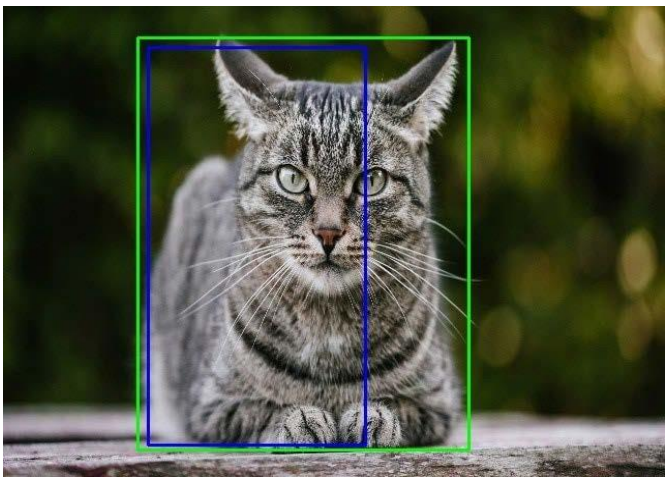

Fig 6.4 Intersection Over Union

In the image above, there are two bounding boxes, one in green and the other one in blue. The blue box is the predicted box while the green box is the real box. YOLO ensures that the two bounding boxes are equal.

Combination of the three techniques

The following image shows how the three techniques are applied to produce the final detection results.

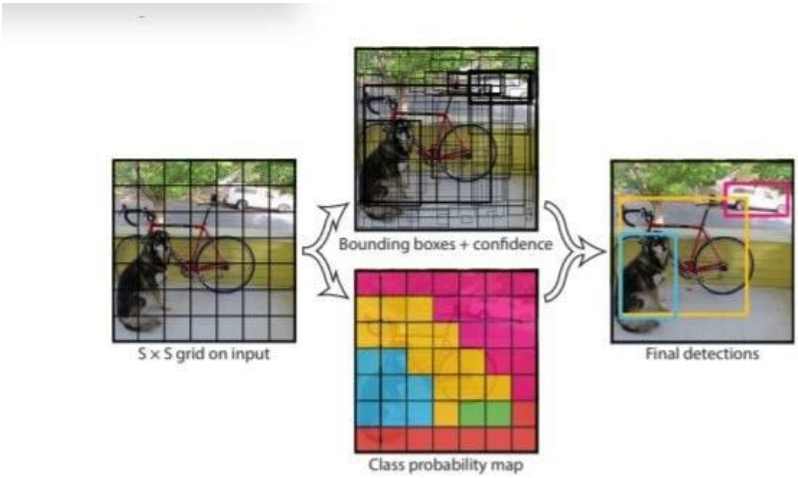

Fig 6.5 Combination of three techniques

First, the image is divided into grid cells. Each grid cell forecasts B bounding boxes and provides their confidence scores. The cells predict the class probabilities toestablish the class of each object.

For example, we can notice at least three classes of objects: a car, a dog, and a bicycle.All the predictions are made simultaneously using a single convolutional neural network.

Intersection over union ensures that the predicted bounding boxes are equal to the realboxes of the objects. This phenomenon eliminates unnecessary bounding boxes that donot meet the characteristics of the objects (like height and width). The final detection will consist of unique bounding boxes that fit the objects perfectly.

For example, the car is surrounded by the pink bounding box while the bicycle is surrounded by the yellow bounding box. The dog has been highlighted using the blue bounding box.

6.4.2 Advantages of YOLO algorithm

- Fast. Good for real-time processing.
- Predictions (object locations and classes) are made from one single network. Can be trained end-to-end to improve accuracy.
- YOLO is more generalized. It outperforms other methods when generalizing from natural images to other domains like artwork.

6.4.3 Disadvantages of YOLO algorithm

The YOLOv3 AP does indicate a trade-off between speed and accuracy for using YOLO when compared to Retina Net since Retina Net training time is greater than YOLOv3.However, the accuracy of detecting objects with YOLOv3 can be made equal to the accuracy when using Retina Net by having a larger dataset, making it an ideal option for models that can be trained with large datasets. An example of this would be common detection models like traffic detection, where plenty of data can be used to train the model since the number of images of different vehicles is plentiful. On the other hand, YOLOv3 may not be ideal to use with niche models where large datasets can be hard to obtain.

6.5 Overall View On Basis Of Implementation

1. Non maxima Suppression
2. NMS Threshold
3. Video_capture.read ()
4. Blob function blob ()
5. Pytesseract ()

Non maxima Suppression

Non Maximum Suppression is a computer vision method that selects a single entity out of many overlapping entities (for example bounding boxes in object detection). The criteria is usually discarding entities that are in a given probability bound.

NMS Threshold

It is defined as the random value given to a system that detects the image on above the particular value otherwise will get undetected.

Video_capture. read()

Function used to read the function that is captured in the video. Or function used to take the captured video as input to the system.

Blob function blob ()

The Blob object represents a blob, which is a file-like object of immutable, raw data they can be read as text or binary data, or converted into a Readable Stream so its methods can be used for processing the data.

Pytesseract ()

Python-tesseract is a wrapper for Google's Tesseract-OCR Engine. It is also useful as a stand-alone invocation script to tesseract, as it can read all image types supported by the Pillow and Laptronica imaging libraries, including jpeg, png, gif, bmp, tiff, and others. Additionally, if used as a script, Python-tesseract will print the recognized text instead of writing it to a file.

CHAPTER 7

SNAPSHOTS AND RESULTS

Fig 8.1 The given video input is showing the text

The output of the video i n the above figure, we report that text recognition result, with a processing speed of 1iterations/second and language is English.

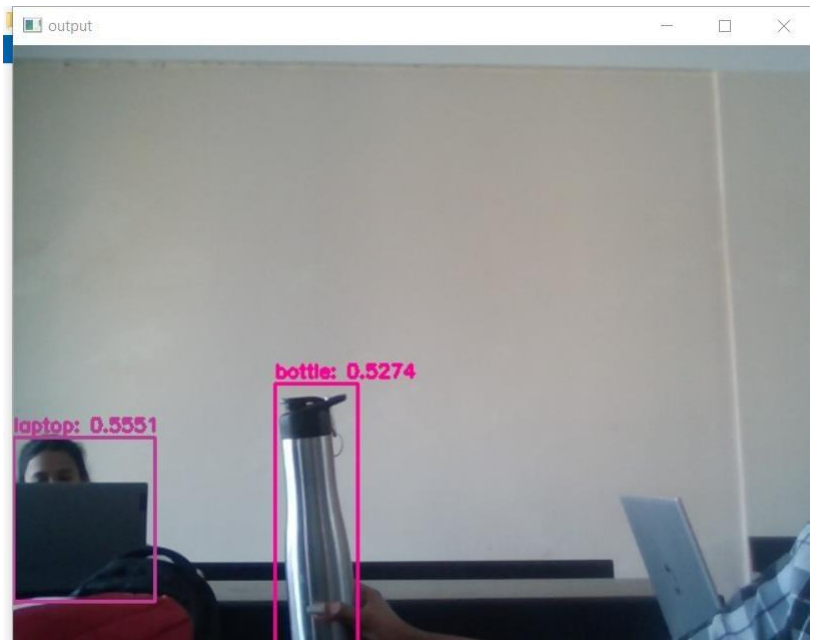

Fig 8.2 The given video input is showing the object

The output of the video in the above figure shows, the object detection using YOLO algorithm with accuracy.

CONCLUSION AND FUTURE SCOPE

This project helps a real-world experience through a Speech Based Feedback System for Visually Impaired People. It solves the challenges and aids visually impaired people to get their tasks done in the same manner as that of a normal person. Therefore, is aiming to make the living of visually impaired people easier as well as help them get through their daily activities without coming in contact with any dangerous obstacles as without it visually impaired people would be living a restricted life.

The future perspective is to increase the object recognition rate which can be achieved by using the TensorFlow library and to provide an exact distance measurement between the people and object. However, for developing an application thatinvolves many objects that are fast-moving, you should instead consider faster hardware. Further, we can implement face recognition and text recognition in the same system. Thus, making the system compatible overall. Future work will explore options to run multiple semantic image segmentation models simultaneously at a higher inference rate. At the time of this work, the OAK-D sensor was a Kickstarter project due to which we were unable to obtain multiple sensors. The plan to evaluate system performance with multiple OAK-D sensors simultaneously. They will also experiment with traditional point cloud methods to detect elevation changes using the Generation2 Depth AI module and incorporate robust object tracking across frames for more accurate traffic analysis. More importantly, based on the insights gained.

REFERENCES

[1] Leo Abraham, Nikita Sara Mathew, Liza George, Shebin Sam Sajan ,'' **Speech Based Feedback System for the Visually Impaired**",in Proceedings of the FourthInternational Conference on Trends in Electronics and Informatics (ICOEI 2020)IEEE Xplore Part Number: CFP20J32-ART; ISBN: 978-1-7281-55180.

[2] Deven Pawar, Mihir Raul, Pranav Raut, Sharmila Gaikwa**," Recognize Objects for Visually Impaired using Computer Vision** ",in International Journal of RecentTechnology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-6, March 2020.

[3] Md. Atikur Rahman, Muhammad Sheikh Sadi**," IoT Enabled Automated Object Recognition for the Visually Impaired**", in Computer Methods and Programs in Biomedicine UpdateVolume 1 ISSN 2666-9900.

[4] Hanen Jabnoun, Benzarti Faouzi, Hamid Amiri," **Object detection and identification for blind people in video scene**", in LR-11-ES17 Signal, Images et Technologies de l'Information (LR-SITI-ENIT) Université de Tunis El Manar, Ecole Nationale d'Ingénieur de Tunis 1002, Tunis Le Belvédère, Tunisie, Conference Paper · December 2015.

[5] Ali Jasim Ramadhan,"**Wearable Smart System for Visually Impaired People**", in Department of Computer Techniques Engineering, Alkafeel University College, Kufa 31003,Published: 13 March 2018.

[6] Preetjot Kaur, Roopali Garg," **Camera & Sensors-Based Assistive Devices For Visually Impaired Persons: A Systematic Review**", INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 8, ISSUE 08, AUGUST 2019.

[7] Ramiro Velazquez, "**Wearable Assistive Devices for the Blind**", LNEE 75, Springer, pp 331-349, 2010.

[8] Apoorva Raghunandan, Mohana Mohana, Raghav Pakala," **Object Detection Algorithms for Video Surveillance Applications**", Conference Paper · April 2018.