Project Report on

Object Detection

Submitted in partial fulfillment of the

requirement for the award of the degree of

B.Tech, Computer Science & Engineering

Under The Supervision of
Dr. AJAY SHANKER
Professor

Submitted By

Mohammad Hussain Zaidi(18SCSE1010075)
Ahmad Yasir(18SCSE1010586)

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
May, 2022

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING, GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

We hereby certify that the work which is being presented in the project, entitled **"Object Detection "** in partial fulfillment of the requirements for the award of the B.TECH submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of a month, Year to Month and Year, under the supervision of Dr. Ajay Shanker Singh, Professor, Department of Computer Science and Engineering, of School of Computing Science and Engineering, Galgotias University, Greater Noida

The matter presented in the project has not been submitted by us for the award of any other degree of this or any other places.

Mohammad Hussain Zaidi(18SCSE1010075)

Ahmad Yasir(18SCSE1010586)

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Dr. Ajay Shanker Singh
Professor

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of Mohammad Hussain Zaidi(18SCSE1010075) & Ahmad Yasir(18SCSE1010586) has been held on May 2022 and his/her work is recommended for the award of B.Tech.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: May, 2022
Place: Greater Noida

ACKNOWLEDGEMENT

We would like to express our special thanks of gratitude to my guide Dr. Ajay Shanker Singh who gave me the golden opportunity to do this wonderful project on the topic Object Detection, which also helped me in doing a lot of Research and I came to know about so many new things we are really thankful to them.

Thank you.

Abstract

The issue of visual impairment or blindness is faced worldwide. According to statistics of the World Health Organization (WHO), globally, at least 2.2 billion people have a vision impairment or blindness, of whom at least 1 billion are blind. In terms of regional differences, the prevalence of vision impairment in low- and middle-income regions is four times higher than in high-income regions. Blind people generally have to rely on white canes, guide dogs, screen-reading software, magnifiers, and glasses to assist them for mobility, however, to help the blind people the visual world has to be transformed into the audio world with the potential to inform them about objects as well as their spatial locations. Therefore, we propose to aid the visually impaired by introducing a system that is most feasible, compact, and cost effective.

The goal of the project is to develop a model an object recognition model to detect and recognize the object for visually impaired people and it can also be used in other markets. Existing techniques for object detection requires a large training data it takes more time and the process is complicated. In this project we propose computer vision concept to recognize object and convert it into text by using the pre trained dataset and the texts are converted further to speech. The system will also support multiple object detection and building a one line description of the object which can be either describe the position of the object or the color. The proposed system can further be implemented in any portable devices for the visually impaired people to recognize objects.

Keywords— Visual Impairment, Computer Vision, Object Recognition, voice output.

Contents

List of Figures

CHAPTER-1 Introduction

"ONLY BECAUSE ONE LACKS THE USE OF THEIR EYES DOES NOT MEAN THAT ONE LACKS VISION." Eyesight is one of the essential human senses, and it plays a significant role in human perception about the surrounding environment. For visually impaired people to be able to provide, experience their vision, imagination mobility is necessary. The International Classification of Diseases 11 (2018) classifies vision impairment into two groups, distance and near presenting vision impairment. Globally, the leading causes of vision impairment are uncorrected refractive errors, cataract, age-related macular degeneration, glaucoma, diabetic retinopathy, corneal opacity, trachoma, and eye injuries. It limits visually impaired ability to navigate, perform everyday tasks, and affect their quality of life and ability to interact with the surrounding world upon unaided. With the advancement in technologies, diverse solutions have been introduced such, as the Eye-ring project, the text recognition system, the hand gesture, and face recognition system, etc. However, these solutions have disadvantages such as heavyweight, expensive, less robustness, low acceptance, etc. hence, advanced techniques must evolve to help them. So, we propose a system built on the breakthrough of image processing and machine learning. The proposed system captures real-time images, then images are pre-processed, their background and foreground are separated. The extracted features are matched with known object features to identify the objects. Once the object is successfully recognized, the object name is stated as voice output with the help of text-to-speech conversion .We want to explore the possibility of using the hearing sense to understand visual objects. The sense of sight and hearing sense share a striking similarity: both visual object and audio sound can be spatially localized. It is not often realized by many people that we are capable at identifying the spatial location of a sound source just by hearing it with two ears.

In this project we propose computer vision concept to recognize object and convert it into text by using the pre trained dataset and the texts are converted further to speech. The system will also support multiple object detection and building a one line description of the object which can be either describe the position of the object or the color. The proposed system can further be implemented in any portable devices for the visually impaired people to recognize objects.

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 suffers 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 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. In this project, we want to explore the possibility of using the hearing sense to understand visual objects. The sense of sight and hearing sense share a striking similarity: both visual object and audio sound can be spatially localized. It is not often realized by many people that we are capable at identifying the spatial location of a sound source just by hearing it with two ears. In our project, we build a real-time object detection and position estimation pipeline, with the goal of informing the user about surrounding object and their spatial position using binaural sound.

According to the World Health Organization, there are approximately 285 million people who are visual impairments, 39 million of them are blind and 246 million have a decrease of 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.

In the computer vision community, developing visual aids for handicapped persons is one of the most active research projects. Mobility aids are intended to describe the environment close to the person with an appreciation of the surrounding objects. These aids are essential for fine navigation in an environment described in a coordinate system relative to the user. In this paper, we present an overview of vision substitution modalities [1-12] and their functionalities. Then, we introduce our proposed system and the experiments tests.

1.1 Formulation of Problem:

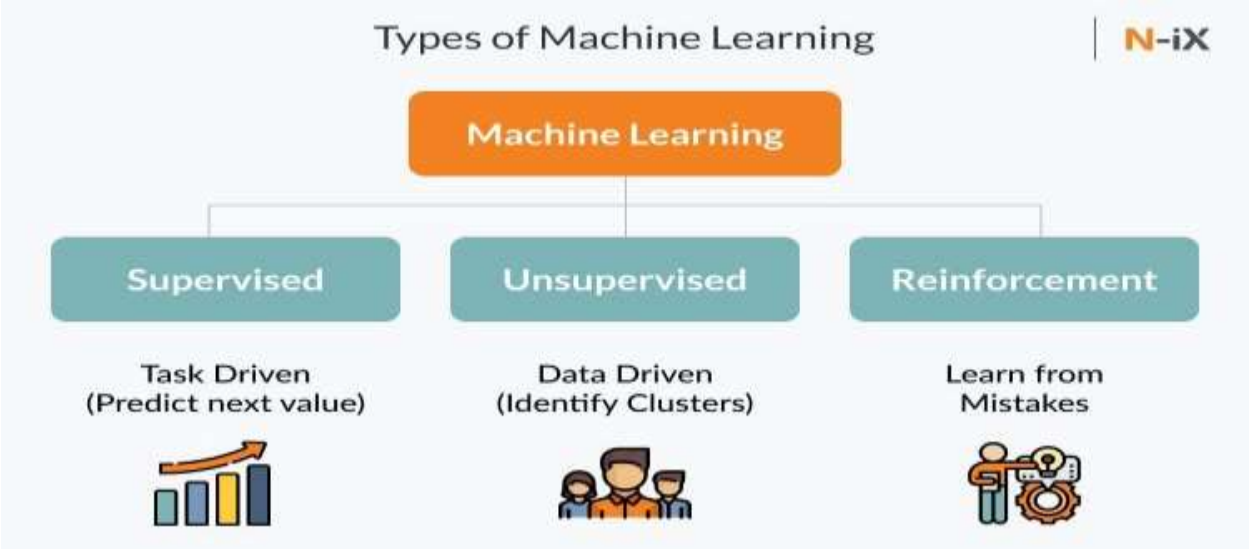
In this project we propose computer vision concept to recognise object and convert it into text by using the pre trained dataset and the texts are converted further to speech. The system will also support multiple object detection and building a one line description of the object which can be either describe the position of the object or the colour. The proposed system can further be implemented in any portable devices for the visually impaired people to recognise objects. The goal of the project is to develop a model an object recognition model to detect and recognise the object for visually impaired people and it can also be used in other markets. Existing techniques for object detection requires a large training data it takes more time and the process is complicated.

1.2.1 Tool and Technology Used:

Machine Learning Analytical model building is computerized by the method of data examination in Machine learning. It deals with the department related to artificial intelligence hinge on the machine, can understand the data, detect design, and also make resolutions with even human involvement. These machine learning charges are broadly divided into numerous divisions.

Machine learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

Analytical model building is computerized by the method of data examination in Machine learning. It deals with the department related to artificial intelligence hinge on the machine, can understand the data, detect design, and also make resolutions with even human involvement.23 These machine learning charges are broadly divided into numerous divisions, as shown in Figure 1, namely supervised learning and unsupervised learning



Supervised Model: These algorithms were instructed using labelled instances, such as retrieving the required output by giving specified inputs. An example, such as detecting the online marketing of products, the outcome would be either "order delivered successfully" or "not delivered," in which it is mentioned in the system language of Boolean values either true or false. This learning algorithm retrieves accurate outputs proportional to the group of inputs they receive. This algorithm also identifies the exact outputs to the original outputs on collations for the detection of the errors, and it alters the algorithm consistently. There are also supervised learning techniques that use specific designs to forecast the values of the label on additional unlabeled data, namely classification, regression,

prediction, and gradient boosting. In an application such as foretelling historical data like upcoming events, this supervised learning is frequently used. For instance, it precedes in deceitful cases like credit card transactions or when an insurance customer is at a point in filing a claim. The Regression algorithms include any value within a specific collection that produces uninterrupted outputs. For example, uninterrupted outputs are the values like temperature, length, or price of the object.

Unsupervised Model: These algorithms develop an algebraic model form the group of data, which takes only inputs irrespective of the output flags. The assembling or congregation of data tips to attain a beautiful structure in the data obtained through this unsupervised learning algorithm. As in attribute learning, inputs grouped into divisions by collecting patterns in the data can also be achieved through unsupervised learning. The procedure involving reducing the count of "features," or the inputs among the data set, is often termed to as dimensionality reduction.

Another learning mechanism that uses identical implementations as supervised learning is referred to as Semi-supervised learning. Priming can be handled in labelled and unlabelled data – Expendently smaller amounts of labelled data with many unlabelled data (Unlabelled data is cheaper and essential for lesser endeavours to gain). Classification, regression, and prediction methods are implemented in this part of learning. Semisupervised learning is applicable or maintained to its fully labelled training process only when the cost is related to it too high. An example of this type of knowledge is detecting an individual's face in front of a web camera. The significant role in monitoring the patient's health in health care instantaneously by providing wearable devices and sensors in this vast developing health industry. It is also used in diagnosis and treatment methods to detect the courses or errors in the medical expert's analysis. In government sectors, vast data is accessed through different sources that can be detected for the perception that needs to be prevented and made safer by the tools that mainly need machine learning technology. Discovering the confidential data, for instance, finds ways to increase its accuracy and save money. Identifying the deception and reducing detection frauds can also be maintained through machine learning.

Object detection is an important computer vision task used to detect instances of visual objects of certain classes (for example, humans, animals, cars, or buildings) in digital images such as photos or video frames. The goal of object detection is to develop computational models that provide the most fundamental information needed by computer vision applications: “What objects are where?”.

- Detecting a specified object class within an image.
- Object detection has many applications in computer based vision.
- Object tracking, object recognition, scene surveillance.
- The focus of this project was to implement object detection, and to detect objects. Object detection refers to the capability of computer and software systems to locate objects in an image/scene and identify each object. Object detection has been widely used for face detection, vehicle detection, pedestrian counting, web images, security systems and driverless cars.

Based on the lightening conditions, some of the objects, particularly in the indoor case, the image might not be visualized depending upon the pixels that have been detected by the system.12 In lightening conditions, constituting the captured image input, which is not recognized by the system, has become the most challenging task. Putting these issues under consideration, the work has become very challenging and dense. The potential of traversing from one location to another is that which is related to our day to day lives. For a computer to understand circumstances to identify and discover the objects in images and videos plays a significant task. As for humans, this is very simple when compared to that of the system to be recognized. In the visually impaired persons, this has to become a challenging task in solving the discrimination. The object detection algorithm needs to take out the features pertained to certain specific classes; there might be a large quantity of pre- interpretation is required. Detecting and recognizing the object particular to the indoor captured image, which is taken as an input image by the system, should also detect the adjoining environments in its surroundings.

Visual Studio Code is a source-code editor made by Microsoft for Windows, Linux and mac OS. Features include support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git Users can change the theme, keyboard shortcuts, preferences, and install extensions that add additional functionality. Visual Studio Code is a source-code editor that can be used with a variety of programming languages, including Java, JavaScript, Go, Node.js, Python and C++.It is based on the Electron framework, which is used to develop Node.js Web applications that run on the Blink layout engine. Visual Studio Code employs the same editor component (codenamed "Monaco") used in Azure DevOps (formerly called Visual Studio Online and Visual Studio Team Services).

CHAPTER-2

Literature Survey

Related Work

There exist multiple tools to use computer vision technologies to assist blind people. The mobile app Tap Tap See uses computer vision and 1 crowdsourcing to describe a picture captured by blind users in about 10 seconds. The Blindsight offers a mobile app Text Detective featuring optical character recognition (OCR) technology to detect and read text from pictures captured from the camera. Facebook is developing image captioning technology to help blind users engaging in conversations with other users about pictures. Baidu recently released a demo video of a DuLight project. No further details of the product is available at the moment. However, the product video suggests concepts of describing scenes and recognizing people, money bills, merchandises, and crosswalk signal. However, these products were not focusing on enabling general visual sense for blind people and did not use the spatial sound techniques to further enhance the user experience. Some works exist in the general scope of sensory substitution. Daniel Kish, who are totally blind, developed accurate echolocation ability using “mouth clicks” for navigation tasks including biking and hiking independently. Colorblind artist Neil Harbisson developed a device to transform color information into sound frequencies. An extreme attempt of converting visual sense to sound is introduced by the voice technology. The vOICe system scans each camera snapshot from left to right, while associating height with pitch and brightness with loudness. However, all these attempts on sensory substitution are reported with very difficult learning process. In contrast, we utilize visual recognition algorithms which lead to more direct ways of understanding objects from a visual scene. The use of 3D sound technology for providing useful information and assisting blind people has also been investigated by researchers. introduced a system that uses spatial audio to facilitate discovery of points of interest in large, unfamiliar indoor environments (e.g. shopping mall). tries to integrate 3D sound into GPS-based outdoor navigation product. However, no visual recognition has been used in those works. The use of object detection techniques can open up new possibilities in assisting indoor navigation for blind and visually impaired people.

Related works show that visual substitution devices accept input from the user's surroundings, decipher it to extract information about entities in the user's environment, and then transmit that information to the subject via auditory or tactile means or some combination of these two. Among the various technologies used for blind people, the majority is aids of mobility and obstacle detection . They are based on rules for converting images into data sensory substitution tactile or auditory stimuli. These systems are efficient for mobility and localization of objects which is sometimes with a lower precision. However, one of the greatest difficulties of blind people is the identification of their environment and its . Indeed, they can only be used to recognize simple patterns and cannot be used as tools of substitution in natural environments. Also, they don't identify objects (e.g. whether it is a table or chair) and they have in some cases a late detection of small objects. In addition, some of them seek additional auditory, others require a sufficiently long period for learning and testing. Among the problems in object identification, we note the redundancy of objects under different conditions: the change of viewpoint, the change of illumination and the change of size. We have the concept of intra-class variability (e.g. there are many types of chairs) and the inter-class similarity (e.g. television and computer). For this reason, we are interested in the evaluation of an algorithm for fast and robust computer vision application to recognize and locate objects in a video scene. Thus, it is important to design a system based on the recognition and detection of objects to meet the major challenges of the blind in three main categories of needs: displacement, orientation and object identification.

In the fields of Image Processing and Object Recognition a lot of exploration is being performed. Researchers have tried to develop more and more flexible, compact and reliable Object Recognition system for visually impaired people in the respective fields. However, all of them used various methods to make these Object Recognition systems successful and reliable for daily use. Recognizer is an iOS application by LookTel. It is a commercial application that is supposed to detect an object within the camera boundary view that was previously stored in a local database of objects or images. The application is delivered to help visually impaired user to recognize household objects. For reliable results, objects

stored in the database should be captured by a sighted person in a predefined orientation. But, the limitation of the tool is that, as, the image is pre-loaded by a sighted person could be a drawback. Cause every time the blind user has to be dependent on the sighted person for its reliability. An additional application projected for the blind users is the EyeRing project. The EyeRing project consists of a VGA camera, AVR microcontroller, a Bluetooth module with control keys. This is a finger-worn device, converses with an Android smart phone. The chore of smart phone is dealing with speech processing algorithms and all image processing algorithms. This solution, however, is expensive for the visually impaired people. As, this project is equipped with many of the instruments, makes this project complex to carry for a visually impaired person.

Mouna and Riadh have proposed their work on "An Evaluation of Retina Net on Indoor Object Detection for Blind and Outwardly Impaired Persons Assistance Navigation" in 2020.In this work, the capacity of PC vision is to distinguish indoor items precisely. The outwardly hindered individuals can be helped by exploring the reasons for the CNN framework.4,5,14 To recognize the particular items first, we really want to distinguish the pixels accessible in the pictures. In case the lighting conditions aren't right, then, at that point, it is trying to catch and distinguish the articles with high exactness. To identify the indoor articles,the calculation needs to extricate the picture highlights with a specific class, and it tends to be finished by RetinaNet.25 To empower the organization for little article identification by a Region Proposal Networks (RPN), which includes subsampling to acquire the picture data. The Resort with 152 examples accomplished a normal accuracy with 83.1%, and DenseNet with 121 tests accomplished a normal accuracy with 79.8%. Han Hu and Jiayuan have proposed their work on "Connection Networks for Object Detection" in 2018. In view of the connection models, this work relegated an equivalent amount of work by thinking about its highlights. This eliminates duplication and achieves precision at explicit guidelines. Since the items are adjusted in the 2D scale proportion, it utilizes protests rather than words. Further, the model is ordered into two parts that fall under mathematical and unique weights.

Xiangrong and Alan have proposed their work on "A TimeEfficient Cascade for Real-Time Object Detection: With applications for the outwardly disabled" in 2005. In this work, the primary target is to zero in principally on time intricacies and their exactnesses relying on the different test that has been performed by the covetous methodology the module which distinguishes the text in the pictures which can be improved for outwardly disabled individuals. The nature of the model can be estimated by F.P. furthermore F.N. rates. The choice capacity of the calculation should be possible by a bunch of preparing pictures and classifiers. The brilliant adaptive framework will be utilized for vision issue individuals. On the miniature screen visuals, the picture addresses itself in a stress way leaving specific spots of the picture behind.

Alice Tang and Zhiyuan have proposed their work on "Programmed Registration of Serial Cerebral Angiography: A Comparative Review" in 2018. During times, in view of this work basically in the clinical field, explicit changes have been made in recognizing the infection and recuperations by thinking about its and exactness. Attractive reverberation imaging (X-ray) and figured tomography (C.T). are investigated on picture handling calculations that are exceptionally analyzed rather than DSA. While DSA is alluded to the determination of a few neurovascular conditions which is utilized at the hour of medical procedures, on these contemplations, it very well may be inferred that the structure is planned dependent on the patients analyzed with ischemic stroke.

Wei and Xia have proposed their work on "HCP: A Flexible Convolutional Neural Network (CNN) Framework for Multi-mark Image Classification" in 2015. In this work, a CNN model delivers the best presentation for picture order with a solitary name. Because of intricacy, multi naming is an open test for preparing picture designs. A solitary picture object is taken as an info will be given for speculations extraction, and this is imparted to CNN to get individual scores by max pooling. The picture's speculations are related to various shadings that can be shown by various clusters.The extraction technique produces prescient outcomes that are used by max pooling. By contrasting the I-FT and HCP models,

the HCP model further develops the framework execution by 5.7%.Edge and Issam have proposed their work on "Indoor Object Acknowledgment in a mix of a RGB(Red-GreenBlue) picture and its comparing profundity picture (RGBD) Images with Complex-Valued Neural Networks for Visually-Impaired People" in 2018. In this work, the multi-model is utilized for outwardly disabled individuals to recognize the items with a multi-class system in an indoor region. This model requires some investment more than one name. The CVNN and multi-name methods partner the picture with names that compare to classes of items at once.16 The groups can be made dependent on multi marking by ML-CVNN, and the LCVNN technique works by picture change to group the issue by positioning arrangement. The info methodology catches the picture by multi-name what's more multi classes to produce the settings of sensible and non-practical of settled and selective structures.

Liang and Miachel have proposed their work on "Utilizing multi-mark characterization for acoustic example recognition and helping bird species overviews" in 2016. This model is utilized for recognizing the examples in metropolitan regions like public roads, pouring, cafés, etc.13 This technique describes the sound bites, which yields the examples. The primary impediment of this model is to require a prepared informational index.
In 2012, the proposition of recognizing the articles in the present day with informational indexes containing 1 million casings utilizing a camera had as of now been started by Pirsiavash and D. Ramanan on "Identifying exercises of every day living in First-individual Camera Sees." It should be possible by the planning of a bunch of investigated outlines with objects. Each article can tracked down the most extreme esteem by separating the crude scores. This model does a crossvalidation cycle for the two tests and prepared information to distinguish also eliminate the copies objects. To accomplish precision for object identification disarray grid is utilized for assessing the classifier blunders. The prearranged informational index contains 24 classes of articles with 1200 labels.

Mohamed and Farid have proposed their work on "A Compressive Sensing Approach to Describe Indoor Scenes for Blind People" in 2015. In this work, the items can be recognized by tie individuals through a camera in various indoor conditions. This model

works by multi marking methodology to quantify the Euclidean distance and Gaussian Process. It checks the presence of different articles concerning the information set. Assuming that the presence of specific items is found, then, at that point, it too sees the positions.

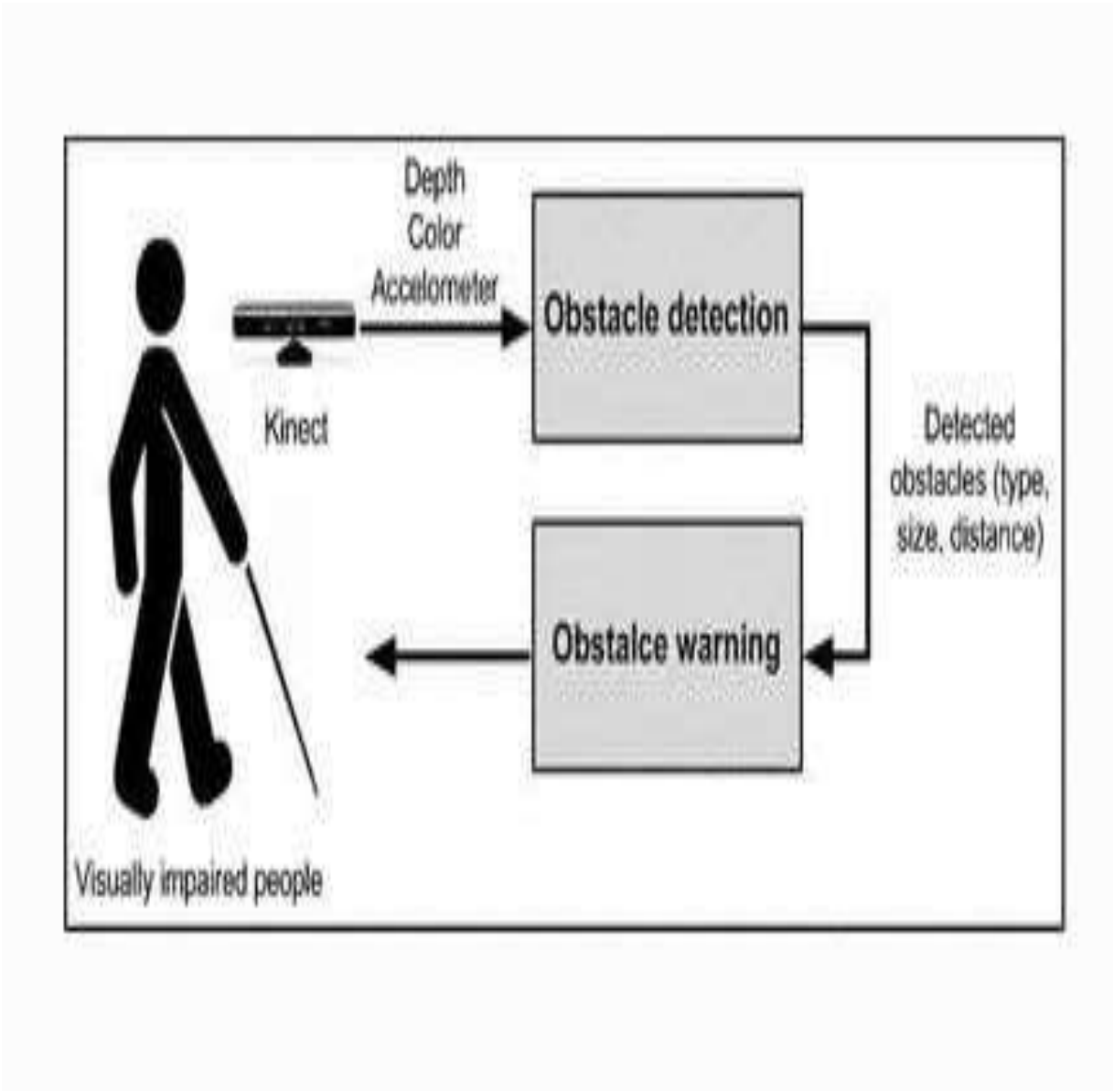
Yong Lee and Ghosh have proposed their work on "Finding notable individuals and items for egocentric video outline" in 2012. In this work, the Camera wearer's day is a thick storyboard instructions the suggested strategies. Then again, in conventional fundamental piece determination methods, the last show of these strategies predominantly looks at the indispensable articles and individuals who collaborate utilizing this camera wearer. A couple of throws/information parcels needed for the storyboard are reflected by the crucial article driven circum-positions in this technique. In light of our practices 17 hours of egotistical information relying on the current procedures, it shows greatness in saliency and summarisation. This has been done in 4 primary advances; they are:

(a) the picture about a celebrity or item could be anticipated utilizing a book self-driven saliency prompts which it prepares a gathering autonomous relapse model.
(b) Separation of each undertaking/occasion in isolating the video into subcategories of assignments.
(c) Gaining the significance of every occasion by empowering the relapse component.

CHAPTER –3

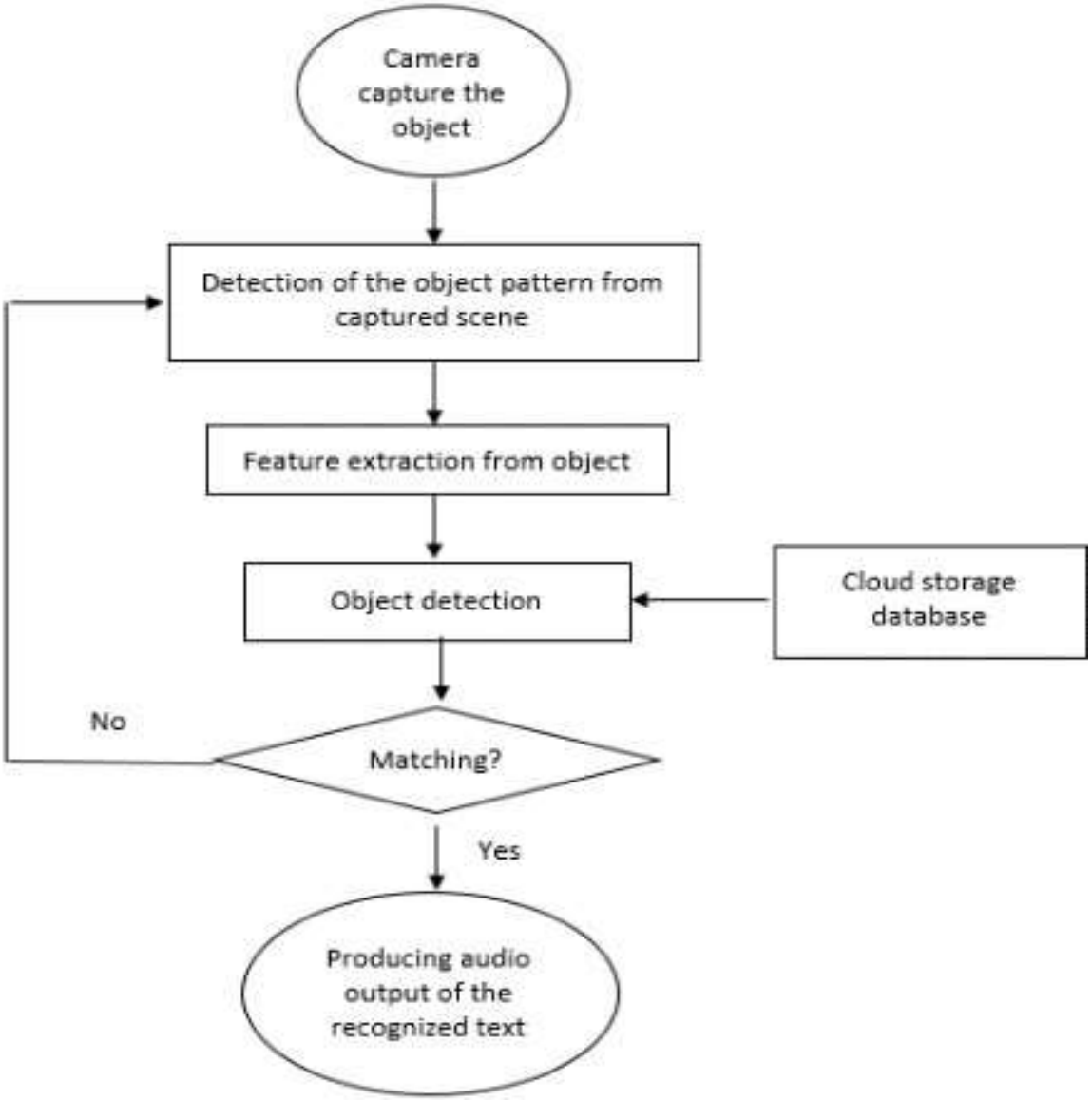
Working of Project

System Flow Chart



Project design

1. Camera capture the object
2. Detection of object pattern
3. Feature extraction from object
4. Object Detection
5. Producing audio output of the recognized text.



We are going to use a model from the Tensorflow Hub library, which has multiple ready to deploy models trained in all kinds of datasets and to solve all kinds of problems. For our use, I filtered models trained for object detection tasks and models in the TFLite format. This format is usually used for IoT applications, for its small size and faster performance than bigger models.

Raw Code:

```
import tensorflow_hub as hub
import cv2 import numpy
import tensorflow as tf
import pandas as pd

# Carregar modelos detector =
hub.load("https://tfhub.dev/tensorflow/efficientdet/lite2/detection/1") labels =
pd.read_csv('labels.csv',sep=';',index_col='ID') labels = labels['OBJECT (2017
REL.)']

cap = cv2.VideoCapture(0)

width = 512 height
= 512

while(True):
   #Capture frame-by-frame
ret, frame = cap.read()

   #Resize to respect the input_shape     inp
= cv2.resize(frame, (width , height ))
#Convert img to RGB     rgb =
cv2.cvtColor(inp,
cv2.COLOR_BGR2RGB)
```

```
    #Is optional but i recommend (float convertion and convert img to tensor image)
rgb_tensor = tf.convert_to_tensor(rgb, dtype=tf.uint8)

    #Add dims to rgb_tensor      rgb_tensor =
tf.expand_dims(rgb_tensor , 0)

    boxes, scores, classes, num_detections = detector(rgb_tensor)

    pred_labels = classes.numpy().astype('int')[0]

    pred_labels = [labels[i] for i in pred_labels]
pred_boxes = boxes.numpy()[0].astype('int')
pred_scores = scores.numpy()[0]

   #loop throughout the detections and place a box around it        for score,
(ymin,xmin,ymax,xmax), label in zip(pred_scores, pred_boxes, pred_labels):         if
score < 0.5:              continue

        score_txt = f'{100 * round(score,0)}'
        img_boxes = cv2.rectangle(rgb,(xmin, ymax),(xmax, ymin),(0,255,0),1)               font =
cv2.FONT_HERSHEY_SIMPLEX         cv2.putText(img_boxes,label,(xmin, ymax-10), font,
0.5, (255,0,0), 1, cv2.LINE_AA)         cv2.putText(img_boxes,score_txt,(xmax, ymax-10), font,
0.5, (255,0,0), 1, cv2.LINE_AA)  #Display the resulting frame      cv2.imshow('black and
white',img_boxes)      if cv2.waitKey(1) & 0xFF == ord('q'):         break

# When everything done, release the capture
cap.release() cv2.destroyAllWindows()
```

CHAPTER – 4 IMPLEMENTATION

Module Description:

The image or video can be loaded into the object detection model. This interface contains loading an image, running a module to execute the program, the number of detected images detected in the module and play audio for better understanding for visually impaired persons.

The model used to help the outwardly impeded individuals can autonomously perceive objects which are close to them.The respect, joined with the review, perplexed the commencement of these AI calculations for outwardly weakened people in helping the exact route, including indoor and outside conditions.

Object Detection

In view of the easing up conditions, a portion of the articles, especially in the indoor case, the picture probably won't be envisioned relying on the pixels that have been recognized by the framework. In easing up conditions, comprising the caught picture input, which isn't perceived by the framework,has turned into the most difficult assignment. Putting these issues viable, the work has become exceptionally testing what's more thick. The capability of crossing from one area to one more is what is identified with our everyday lives.

For a PC to comprehend conditions to recognize and find the items in pictures and recordings plays a critical task. With respect to people, this is extremely basic when contrasted with that of the framework to be perceived. In the outwardly debilitated people, this needs to turn into a difficult assignment in tackling the separation.

The item recognition calculation needs to take out the highlights related to specific explicit classes; there may be an enormous amount of pre-translation is required. Distinguishing and perceiving the article specific to the indoor caught picture, which is taken as an information picture by the framework, ought to likewise distinguish the connecting conditions in its environmental elements.

How does Object Detection work?

Object recognition is a task of computer vision in the field of machine learning. It is considered to be one of the difficult and stimulating tasks of computer vision.

Object detection or recognition aims to detect all instances of objects of a known class, such as people, cars or faces in an image. Usually, only a small number of instances of the object are present in the image, but there are a very large number of possible locations and scales at which they can occur that need to be explored in a way. or another.

Each detection of the image is signalled with some form of pose information. It’s as simple as the location of the object, a location and a scale, or the extent of the object defined in terms of a bounding box. In some cases, the pose information is more detailed and contains the parameters of a linear or nonlinear transformation.

For example, for face detection in a face detector, the locations of the eyes, nose and mouth can be calculated, in addition to the boundary area of the face.

Computer Vision

For analyzing the Visual world to break and elucidate, which explains computer vision in computer technology. In categorizing the objects' accuracy, machines use deep learning models17 and digital images such as cameras and videos. In the early 1950s, demonstrations have already started in computer vision to identify the keen edges and align the simpler objects with falling under categories such as circles and squares by the techniques of first neural networks. Later in the 1970s, Optional character recognition came into existence of computer vision explicated typed or handwritten data on its primary trading tool. The illustrated data mainly used for the blind as a development.3 In the 1990s, the World Wide Web has evolved, producing sizeable images for examining and various computing facial detection had developed. These evolving text frames supported the analysis of machines in detecting particular persons in pictures and videos.

The image segmentation has to be inspected individually by categorized into various partitions or frames. The object detection indicates detecting a particular object in the image. Upgraded object detection admits multiple objects in a single image. For example, in certain instances, like the football field, an offensive player, a defensive player, a ball, etc. To obtain this X, the Y Coordinate model is implemented for the bounding box and detecting everything inside the region.6,7 The facial recognition technology for Object detection has come up with the latest type that concedes the human face in the entire image and detects as a person in particular. Using pattern detection, a duplication of the shapes, colours, and other visual indicators in the picture. The image classifications are used to bring together into multiple divisions. The feature similarities can be attached to pattern detection that classifies the similitude among the matched objects.

The Staggering evolution of the computer vision includes these advanced effects. Around 50 to 90 %, the accuracy in detecting the objects and dividing them into specific categories has rapidly been increased in less than a decade. In the day-to-day usage of computers, they maintain keen exactness in detecting and reacting to visual inputs compared to humans.

The computer vision can be differentiated in three stages for real automation.

a) Stage 1: On examination 3D technology, which obtains a synchronous view over images and even large sets.2

b) Stage 2: Thousands of labelled or pre-identified images are automated models trained by deep learning models.

c) Stage 3: The ending step is the elucidative step in which the object is detected or arranged.

Tensor Flow

The Computer Vision Python library utilizes a basic ImageAI that urges engineers to consolidate cutting edge Artificial Intelligence elements to its stay alive and temporary applications and frameworks.

In Machine language, the delicate stream goes about as a start to finish open-source stage. The complete, adaptable environment of apparatuses, libraries, and local area assets permits the engineers for the cutting edge in ML and could rapidly assemble and assess ML-fueled applications.

TensorFlow is a free and open-source software library for machine learning and artificial intelligence. It can be used across a range of tasks but has a particular focus on training and inference of deep neural networks. TensorFlow was developed by the Google Brain team for internal Google use in research and production. The initial version was released under the Apache License 2.0 in 2015.Google released the updated version of TensorFlow, named TensorFlow 2.0, in September 2019. TensorFlow can be used in a wide variety of programming languages, most notably Python, as well as Javascript, C++, and Java.This flexibility lends itself to a range of applications in many different sectors.

TensorFlow is Google Brain's second-generation system. Version 1.0.0 was released on February 11, 2017. While the reference implementation runs on single devices, TensorFlow

can run on multiple CPUs and GPUs (with optional CUDA and SYCL extensions for general-purpose computing on graphics processing units).TensorFlow is available on 64-bit Linux, macOS, Windows, and mobile computing platforms including Android and iOS. Its flexible architecture allows for the easy deployment of computation across a variety of platforms (CPUs, GPUs, TPUs), and from desktops to clusters of servers to mobile and edge devices.

TensorFlow computations are expressed as stateful dataflow graphs. The name TensorFlow derives from the operations that such neural networks perform on multidimensional data arrays, which are referred to as tensors. During the Google I/O Conference in June 2016, Jeff Dean stated that 1,500 repositories on GitHub mentioned TensorFlow, of which only 5 were from Google

frozen_inference_graph.pb frozen_inference_graph.pb, is a frozen graph that cannot be trained anymore, it defines the graphdef and is actually a serialized graph and can be loaded with this code:

```
def load_graph(frozen_graph_filename):
    with tf.gfile.GFile(frozen_graph_filename, "rb") as f:
        graph_def = tf.GraphDef()
        graph_def.ParseFromString(f.read())
        return graph_def
tf.import_graph_def(load_graph("frozen_inference_graph.pb"))
```

the saved model is a model generated by tf.saved_model.builder and is has to be imported into a session, this file contains the full graph with all training weights (just like the frozen graph) but here can be trained upon, and this one is not serialized and needs to be loaded by this snippet. The [] are tagconstants which can be read by the saved_model_cli. This model is also often served to predict on, like google ml engine par example:

```
with tf.Session() as sess:
```

```
tf.saved_model.loader.load(sess, [], "foldername to saved_model.pb, only folder")
```

model.ckpt files are checkpoints, generated during training, this is used to resume training or to have a back up when something goes wrong after along training. If you have a saved model and a frozen graph, then you can ignore this.

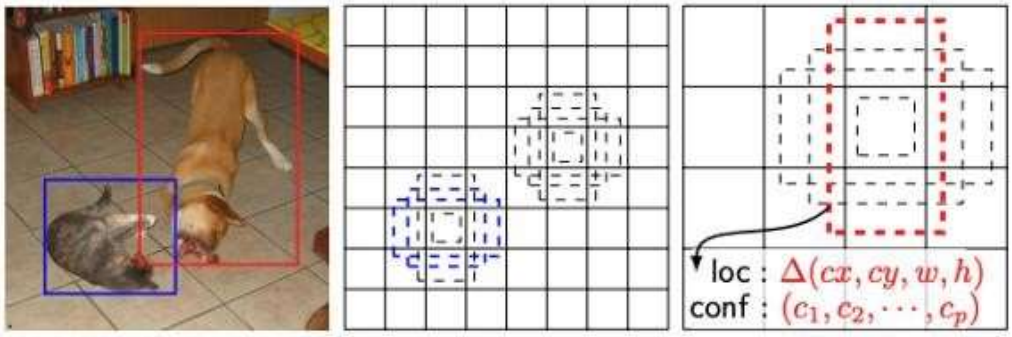
.pbtxt files are basically the same as previous discussed models, but then human readable, not binary. These can be ignored as well.

To answer your conversion question: saved models can be transformed into a frozen graph and vice versa, although a saved_model extracted from a frozen graph is also no trainable, but the way it is stored is in saved model format. Checkpoints can be read in and loaded into a session, and there you can build a saved model from them.

MobilenetSSD

MobilenetSSD is an object detection model that computes the bounding box and category of an object from an input image. This Single Shot Detector (SSD) object detection model uses Mobilenet as backbone and can achieve fast object detection optimized for mobile devices.

MobilenetSSDtakes a (3,300,300) image as input and outputs (1,3000,4) boxes and (1,3000,21) scores. Boxes contains offset values (cx,cy,w,h) from the default box. Scores contains confidence values for the presence of each of the 20 object categories, the value 0 being reserved for the background.



(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

Fig. 1: **SSD framework.** (a) SSD only needs an input image and ground truth boxes for each object during training. In a convolutional fashion, we evaluate a small set (e.g. 4) of default boxes of different aspect ratios at each location in several feature maps with different scales (e.g. 8×8 and 4×4 in (b) and (c)). For each default box, we predict both the shape offsets and the confidences for all object categories ($(c_1, c_2, \cdots, c_p)$). At training time, we first match these default boxes to the ground truth boxes. For example, we have matched two default boxes with the cat and one with the dog, which are treated as positives and the rest as negatives. The model loss is a weighted sum between localization loss (e.g. Smooth L1 [6]) and confidence loss (e.g. Softmax).

In SSD, after extracting the features using an arbitrary backbone, the bounding boxes are calculated at each resolution while reducing the resolution with Extra Feature Layers. MobilenetSSD will concatenate the output of the six levels of resolution and calculate a total of 3000 bounding boxes, and finally, filter out bounding boxes using non-maximum suppression (nms).

Open CV

The open CV remembered for the AI programming library is additionally an open-source PC vision. In business items, it is utilized as an overall framework for PC vision applications and rates up the utilization of machine discernment.

OpenCV (Open Source Computer Vision Library) is a library of programming functions mainly aimed at real-time computer vision. Originally developed by Intel, it was later supported by Willow Garage then Itseez (which was later acquired by Intel). The library is cross-platform and free for use under the open-source Apache 2 License. Starting with 2011, OpenCV features GPU acceleration for real-time operations.

OpenCV is written in C++ and its primary interface is in C++, but it still retains a less comprehensive though extensive older C interface. All of the new developments and algorithms appear in the C++ interface. There are bindings in Python, Java and MATLAB/OCTAVE. The API for these interfaces can be found in the online documentation. Wrappers in several programming languages have been developed to encourage adoption by a wider audience. In version 3.4, JavaScript bindings for a selected subset of OpenCV functions was released as OpenCV.js, to be used for web platforms.

Numpy

NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely. NumPy stands for Numerical Python.

NumPy can be used to perform a wide variety of mathematical operations on arrays. It adds powerful data structures to Python that guarantee efficient calculations with arrays and matrices and it supplies an enormous library of high-level mathematical functions that operate on these arrays and matrices.

Data Set

A Data set is a set or collection of data. This set is normally presented in a tabular pattern. Every column describes a particular variable. And each row corresponds to a given member of the data set, as per the given question.Data sets describe values for each variable for unknown quantities such as height, weight, temperature, volume, etc. of an object or values of random numbers. The values in this set are known as a datum. The data set consists of data of one or more members corresponding to each row.

There are many datasets available that can be used in object detection.

For example:

CIFAR-10:

CIFAR-10 is a comprehensive dataset that consists of 60,000 colour images in 10 different categories. The dataset holds 10,000 test images and 50,000 training images split into five training groups.

Open Images:

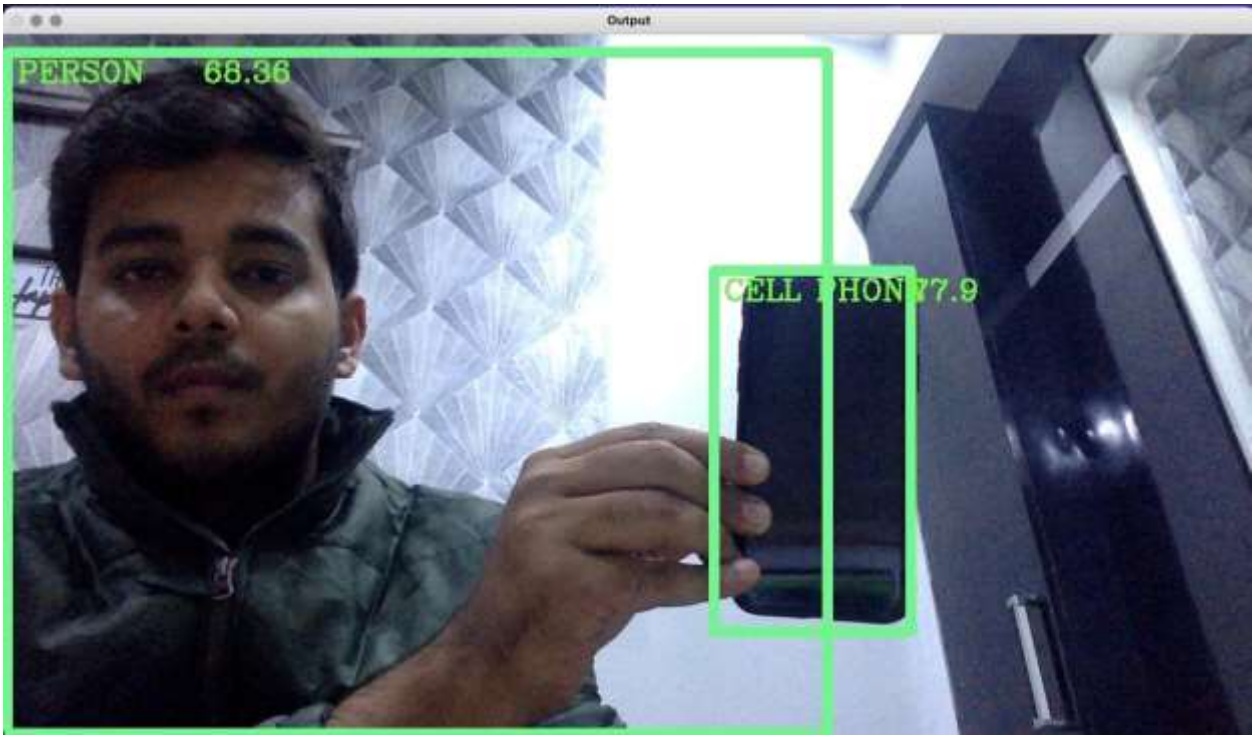
Open Image is a dataset of approximately 9 million pictures annotated with image-level labels, object bounding boxes, object segmentation masks, visual relationships, and localised narratives.

CHAPTER 5
RESULTS AND DISCUSSION

RESULTS

Object Recognition is a process of detecting an object and identifying it using various image algorithms. The main purpose of this paper is to recognize objects in real time and allot the objects to the classes that are previously defined. The algorithms that we utilized are more computationally efficient.

This project detects different types of objects with the help of YOLO pre trained model and uses ffmpeg and gTTS to give the voice feedback. This project works in Real time with the help of webcam or any external camera.



Discussion

The prototype we build successfully recognizes visual objects and presents the detection information as 3D sound, giving the user a sense of "augmented reality". However, the prototype suffers from the following limitations.

First, it is common for user to focus on certain object from afar and navigate to a location close to the object. In this task, the user need a consistent instruction of the target object from approximately 10 m away to only 20 cm away. That impose a very high requirement to the object detection model. To our experience, YOLO can correctly detect objects, such as chair, within a range about 2-5 m away. Objects that are outside this range are either unrecognized or misclassified. One approach to solve this issue is to incorporate training images with greater scale ranges (e.g., include chair picture captured from 20 cm away and 10 m away). However, it may be difficult for object detection models to classify the object from a picture of extreme scale (too close or too far). Another approach to solve this is to use object tracking algorithm to track the object (e.g. a chair) once the user have identified as the target. These two approaches are worth exploring in the future work.

The second issue reported by the blind user is the blocking of ambient sound by using earbuds. However, this can be solved by using bone conduction earphones, which leave ears open for hearing surrounding sounds.

The third issue reported by the blind user is "information overload" when the system is trying to notify user of multiple objects at the same time. This can be solved by delayed notifications. For example, the system can sequentially notify the user of the object from left to right. However this solution requires the user stands still while playing the 3D sounds. Moreover, blind people usually do not want to know every objects in his "eyesight", but instead want to know objects that are pertinent to their immediate need. For example, they may want to find a particular room in a building, or find food and drinks during a conference. In this regard, the system should have three modes: exploration mode where users are notified with every detected objects, search mode where the system only notify users of the object they are looking for, and navigation mode where only 5 the target object and obstacle objects are notified to users in real time. In sum, extensive work is required to analyze users' need if one would like to stem from this prototype to a really helpful assistive product.

CHAPTER-6 Conclusion and Future Scope

Despite the fact that Objection identification is a regarded task yet, it is an imaginative task. It assumes a fundamental part in various executions like distinguishing a picture, autoexplanation of picture, and dread of the belief system. Taking out the issue of vision in outwardly debilitated people, the proposed work can be utilized adequately in recognizing the items alongside their plan designs in a definite way and to distinguish them among numerous various articles in a caught input picture independently with high exactness and with master route, by executing the Specific model X-Y plane by computing their rates precisely of the location and furthermore supporting the change input pictures to discourse. The article recognition likewise outfits its outcomes on different items and different techniques in finding relics, distinguishing and grouping each progression for its efficiency.

This framework could additionally be formed by bringing it into any electronic gadgets and a superior framework can be created to keep away from foundation unsettling influences. Mobilenet design ought to be worked on in future to recognize objects with less certainty level and articles that are more modest with significantly more higher exactness. In this stage of works, we address the recognition of each object in the scene as an individual task; we 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 project that we are trying to develop will advance the way of people that are going to use the computing environment. The application being developed would lead to a better solution for communication and daily living for visually impaired people. It would be a fresh approach in the computing environment, and will lead to a successful and reliable approach. As, we are using all Android based integrations, which is totally an open source, due to which, the product we are developing will be an open source and can be used by every visually impaired person. The application that we are trying to construct is currently

under development and after final completion of the application it can be used for further enhancement for visually impaired people. It can be further used for its use in cybernetics, computing environment and developing systems in digital world on many aspects like for understanding human behaviour based on their way of interaction and for further explores.

References

Authors Jian-DaWuJun-ChingLiu published title “Development of a predictive system for car fuel consumption using an artificial neural network”

Author(s): Enis Gegic, Becir Isakovic, Dino Kečo, Zerina Mašetić, Jasmin Kevrić published in TEM Journal with title “Car Price Prediction using Machine Learning Techniques”

Author(s): Fons Wijnhoven, Olivia Plant published in ICIS 2017 Proceeding with title “Sentiment Analysis and Google Trends Data for Predicting Car Sales”

Author(s): A. Khodayari; A. Ghaffari; M. Nouri; S. Salehinia; F. Alimardani published in IEEE with title “Model Predictive Control system design for car-following behavior in real traffic flow”

Author: Mariana Listiani in Master Thesis Proceeding with title “Support Vector Regression Analysis for Price Prediction in a Car Leasing Application”

Durette, B., Louveton, N., Alleysson, D., and H´erault, J, 2008. Visuoauditory sensory substitution for mobility assistance: testing The VIBE. In Workshop on Computer Vision Applications for the Visually Impaired, Marseille, France.

Hern´andez, A. F. R. et al, 2009. Computer Solutions on Sensory Substitution for Sensory Disabled People. In Proceedings of the 8th WSEAS International Conference on Computational Intelligence, Manmachine Systems and Cybernetics, pp. 134–138.

Tang, H., and Beebe, D. J., 2006. An oral tactile interface for blind navigation. IEEE Trans Neural Syst Rehabil Eng, pp. 116–123.

Auvray, M., Hanneton, S., and O'Regan, J. K., 2007. Learning to perceive with a visuo - auditory substitution system: Localisation and object recognition with 'The vOICe'. Perception, pp. 416–430.

Kammoun, S. et al, 2012. Navigation and space perception assistance for the visually impaired: The NAVIG project. In IRBM, Numro spcial ANR TECSAN, 33(2), pp.182–189.

Brian Katz, F.G. et al, 2012. NAVIG: Guidance system for the visually impaired using virtual augmented reality. In Technology and Disability, pp. 163–178.

Abboud, S., Hanassy, S., Levy-Tzedek, S., Maidenbaum, S., and Amedi, A., 2014. EyeMusic: Introducing a "visual" colorful experience for the blind using auditory sensory substitution. In Restorative neurology and neuroscience.

Martnez, B.D.C., Vergara-Villegas, O.O., Snchez, V.G.C., De Jess Ochoa Domnguez, H., and Maynez, L.O., 2011. Visual Perception Substitution by the Auditory Sense. In Beniamino Murgante, Osvaldo Gervasi, Andrs Iglesias, David Taniar, and Bernady O. Apduhan, editors, ICCSA (2), volume 6783 of Lecture Notes in Computer Science, pp. 522–533. Springer.

Jabnoun, H., Benzarti, F., and Amiri, H., 2014. Visual substitution system for blind people based on SIFT description. In International Conference of Soft Computing and Pattern Recognition. Tunsia, pp. 300–305.

Renier, L. et al, 2005. The Ponzo illusion with auditory substitution of vision in sighted and early-blind subjects. In Perception.

Zhang, J., Ong, S. K. and Nee, A. Y. C., 2009. Design and Development of a Navigation Assistance System for Visually Impaired Individuals. In Proceedings of the 3rd International Convention on Rehabilitation Engineering & Assistive Technology, i-CREATe ’09, New York, NY, USA.

Jafri, R., Ali, SA., Arabnia, HR., and Fatima, S., 2013. Computer visionbased object recognition for the visually impaired in an indoors environment: a survey. In The Visual Computer, Springer Berlin Heidelberg, pp. 1–26.

Lowe, G.D., 2004. Distinctive Image Features from Scale-Invariant Keypoints. In Int. J. Comput. Vision, 60(2), pp. 91–110.

Bay, H. , Ess, A., Tuytelaars,T., and Gool, L.V. 2008. SURF: Speeded Up Robust Features. Computer Vision and Image Understanding (CVIU), pp. 346–359.

S. Hare., J., Sina,S., and P. Dupplaw, D. 2011. OpenIMAJ and ImageTerrier: Java libraries and tools for scalable multimedia analysis and indexing of images. In Proceedings of the 19th ACM international conference on Multimedia (MM '11). ACM, New York, NY, USA, 691- 694. DOI=10.1145/2072298.2072421.

Ondřej, C.. and Matas, J. 2005. Matching with PROSAC-progressive sample consensus. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems, pages 91–99, 2015.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580–587, 2014.

Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32(9):1627–1645, 2010

Jizhong Xiao, Kevin Ramdath, Manor Iosilevish, Dharmdeo Sigh, and Anastasis Tsakas. A low cost outdoor assistive navigation system for blind people. In Industrial Electronics and Applications (ICIEA), 2013 8th IEEE Conference on, pages 828–833. IEEE, 2013.