Eyeris: An Aid for Blind Assistance

Kamal Raj T[1], I Amali Akshaya[2], KV Himasai[3], Kiran N[4], Mohammed Reehan Alam[5]

[1] *Associate professor, CSE, RajaRajeswari College of Engineering, Karnataka, India*
[2] *Student, CSE, RajaRajeswari College of Engineering, Karnataka, India*
[3] *Student, CSE, RajaRajeswari College of Engineering, Karnataka, India*
[4] *Student, CSE, RajaRajeswari College of Engineering, Karnataka, India*
[5] *Student, CSE, RajaRajeswari College of Engineering, Karnataka, India*

ABSTRACT

There is little aid for blind assistance, Therefore, it is necessary to put into action a tool that helps them with their daily tasks. There are existing systems and software that provide visual assistance for reading and accessing a few devices, but these systems lack when the disabled person wants to do some basic tasks like identifying the surroundings in front of them such as a person or object. Therefore, very few mechanisms are invented that aid communication between the blind person and the deaf-dumb person. This project is designed to aid and help a blind person or partially impaired eyesight. This system is developed to aid blind persons without a guardian needed. The software and hardware are designed in a way that helps to detect objects, people, and gestures in vision and recognize them. This is a method that implements object detection and person recognition. For communication between dead-dumb and blind people, we use Sign language which is detected and recognized, and the same is notified to the user. The object or sign is transmitted to a blind person in the form of audio. The idea is to make blind people's lives independent and affordable by offering them affordable solutions.

Keyword: - *Real-time Object detection, Face Recognition, Text-to-Speech, Deep Learning, Visually impaired, Gesture-to-voice.*

---

1. INTRODUCTION

Eyeris is a project that is developed to aid visually impaired people. Blind people lead a difficult life and they rely on others to know about their surroundings or things. They cannot experience the world as we do and it is difficult for them to do basic activities in day-to-day life. Most existing systems only provide a way to do activities such as reading and writing [1]. So to help them Eyeris will perform activities to aid and make them independent.

Eyeris uses different modules to achieve the desired system. It has different modules supporting different operations. The input is captured by the camera as live video and every frame is broken down into frames. The deep neural network processes each frame, marking the appropriate enclosing boxes with various aspect ratios. Then, the best-fit box is selected for the object depending on the prediction score of each enclosing box. For Facial Recognition, each box is compared to an increasing set of facial features. If it fails to analyze even one feature then the box is discarded and determined to not be a facial region [2]. If it is found as the face region then, based on the algorithm, the face is compared with the datasets of faces to see if it recognizes the face or not. The result of any object detection, sign language, or gesture detection is translated into audio using Python gTTS (Google Text-To-Speech) for the person to hear.
The object detection model is trained using YOLO (You Only Look Once), which suggests an end-to-end neural network that predicts bounding boxes simultaneously and class probabilities [3]. The model for detecting objects was trained using the COCO dataset which is large-scale object detection, segmentation, and captioning Dataset.

Eyeris is also made to achieve portability and an easy-to-use interface. This program separates itself into major aspects like object detection, face recognition, gesture recognition, and text-to-speech.

2. 2. EXISTING SYSTEM

A paper “Smart Vision System for Blind” in 2014 [4], was demonstrated to support blind people for movement within unfamiliar surroundings. Another paper used different devices to help indoor and outdoor movement with GPS to track the coordinates of the position of the person using the device [5]. The paper “Oculus” focused on facilitating blind people to self-navigate themselves with no assistance from a third person [6]. Also, there are smartphone-based systems to guide the visually impaired with convenient user interfaces.

Development in the area of computer vision as well is made by processing texts from real-time images. Face recognition by comparing with databases also became an appreciated paper at the IEEE paper conference. A portable and wearable device was developed by merging face detection, face recognition, and audio output modules. A first-rate approach for image processing is given by OpenCV to our stated problem. It is stated in a paper that images are classified using OpenCV and AdaBoost algorithm. The paper “Real-Time Hand Gesture Recognition Using Finger Segmentation” proposed a four-step architecture for recognizing hand gestures. Hand detection is done with the background-subtraction technique. The fingers and palm are separated to distinguish and recognize the fingers. The paper states a problem with the hand movements that are not efficiently recognized if the background color in the stream is the same as the skin color. Using machine learning algorithms to recognize gestures tends to perform better. In another paper, to emphasize object detection based on hue, saturation, and color value range (HSV), the OpenCV application was put into place.

As technology progresses, more effective algorithms, methods, and systems are proposed and created for reliable service as well as real-time detection with less amount of lag and the highest level of accuracy.

This paper uses the YOLO (You Only Look Once) method which is effective in developing reliable systems in real-time. Here the COCO dataset is the one utilized for object detection. One of the most widely used large-scale labeled image datasets made accessible to the public is The Common Object in Context (COCO). YOLO is a single-shot detector that processes an image using a fully convolutional neural network (CNN).

These works served as an inspiration and a copious source of direction for our intended solution, and they also identified potential roadblocks to our advancement. Based on these shortcomings, we set out to design a more progressive solution and smart system to assist the blind.

3. WORKING

The process takes place in the following steps:
1. Capturing the image: A camera is used to collect an ongoing stream of images in real-time.
2. Detecting the objects: On the dataset of various objects, a deep-learning YOLO model is trained. A bounding box is drawn around the object at this stage.
3. Informing the user: The user is informed about the object in the visual frame in the form of audio. This step uses text-to-speech recognition [7].
4. Real-time Face recognition: This prototype also has a face recognition feature that helps the user to identify individuals in the dataset and their names will be transmitted as audio.
5. Gesture recognition: The deaf person's gestures will be recognized, and matched with the gesture datasets, and if recognized, the gesture is pronounced.

4. OBJECT DETECTION

Object identification is probably one of computer vision's most promising subfields. It describes the computer system's capacity to find and recognize instances of objects [5]. Drawing bounding boxes around things that are discovered enables us to locate them in a scene. This is how object detection works. One of the earliest effective attempts to address object detection issues came in 2014 with the development of the R-CNN model [8]. This model used region proposal techniques with convolutional neural networks (CNNs) to locate and detect objects in images. The two primary classifications of object detection algorithms are single-shot detectors and two-stage detectors.

4.1 YOLO (You Only Look Once)

Convolutional neural networks (CNN) are included in the YOLO method to recognize objects in real-time. The algorithm requires only a single forward propagation through the aid of a neural network to find objects. The input of an image into the YOLO algorithm, which then employs a deep convolutional neural network to find objects in the picture.

The CNN model's architecture, the foundation of YOLO, is depicted here. Before passing the input image through the convolutional network, the architecture resizes it to 448x448. A 1x1 convolution is employed to lessen the number of channels, and a 3x3 convolution is then used to develop a cuboidal output. The activation function under the hood is ReLU, except for the final layer, which employs a linear activation function. Some extra techniques, consisting of batch normalization and dropout, respectively regularize the model and prevent it from overfitting.

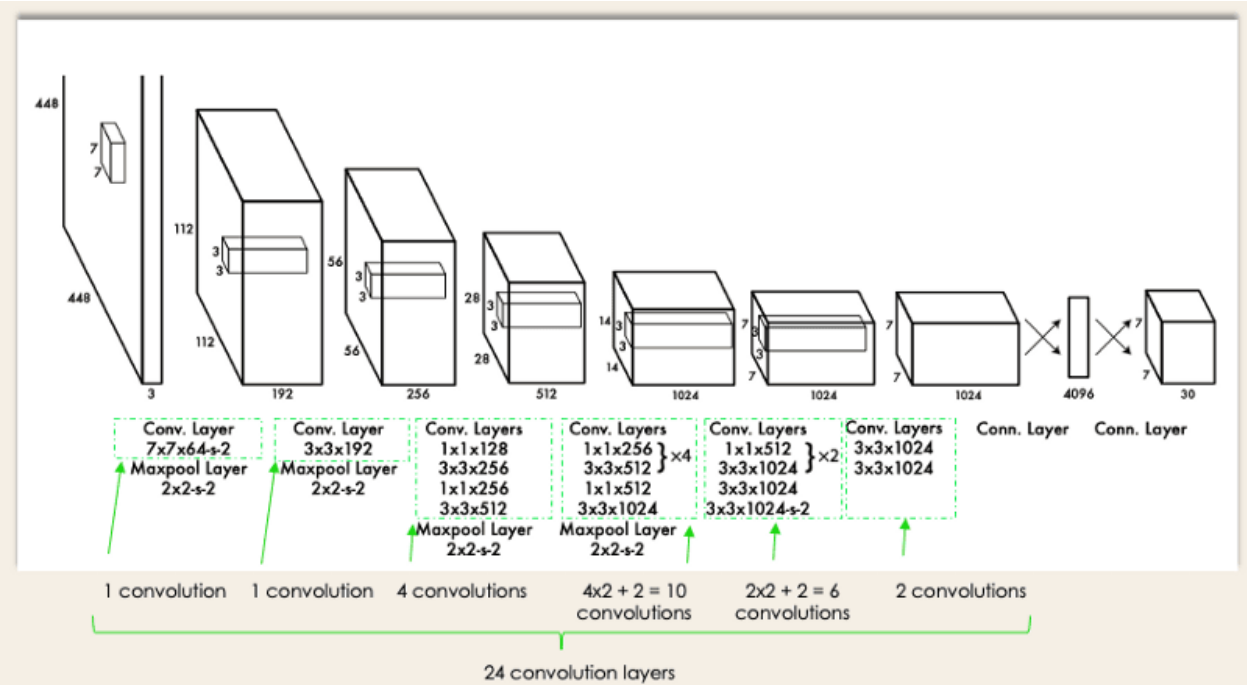

Fig -1: YOLO Architecture Diagram

4.2 COCO Dataset

Among the most well-known large-scale image, datasets are COCO, which stands for Common Object in Context. It features over 1.5 million object instances and image annotations in 80 categories for a sample of the items we encounter every day. The labeling of objects is used to complete instance segmentation which labels every instance of an object in every segmentation [5]. Microsoft created this dataset to facilitate the studies of understanding scenes. Fig. 2 shows some objects of the COCO dataset.

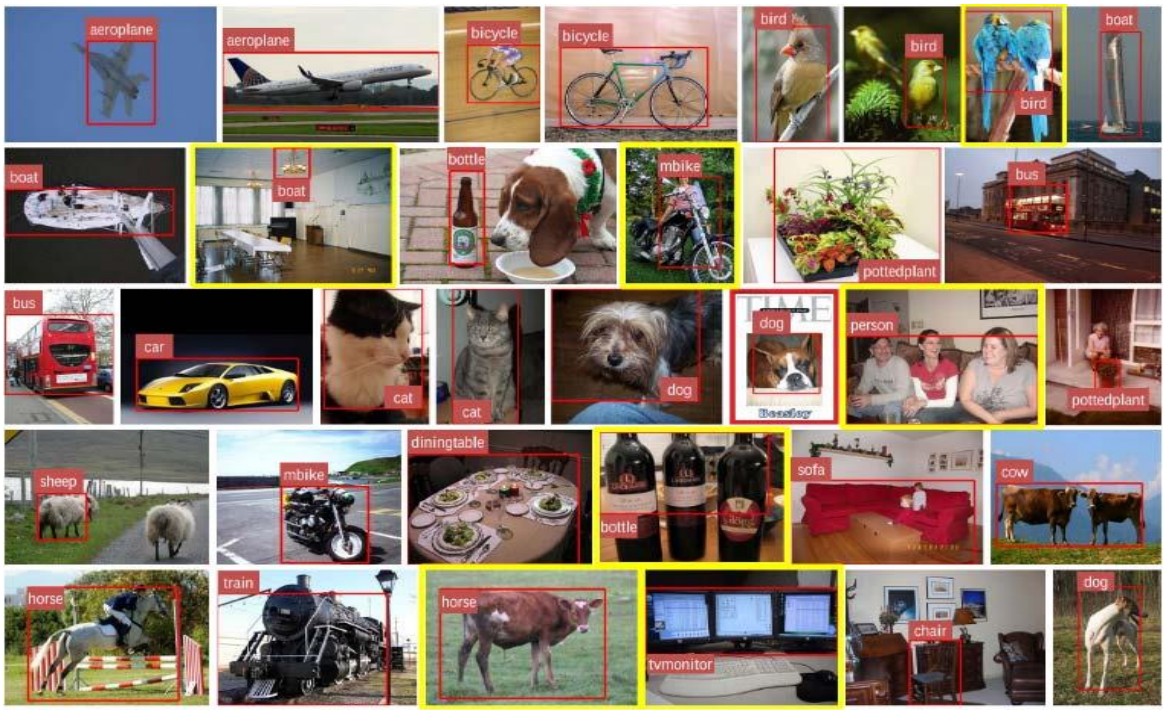

Fig -2: Objects of COCO dataset [5]

Fig. 3 shows the flow of the Object Detection project. A camera records the surrounding scenes and the images are passed to the YOLO model, which finds and recognizes the objects in the images. The recognized object name is presented as text output with bounding boxes around objects. Then the name of the object is passed to the Speech model which converts the object names in the text form to an audio file. Then the audio is played to the user automatically through speakers. So, the user can hear the object's name and understand what object is exactly in the scene.

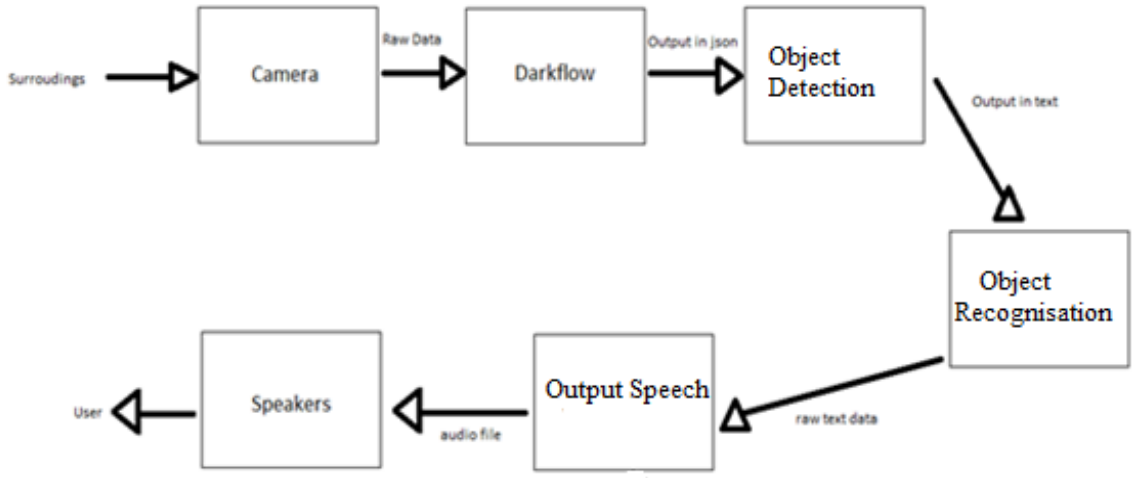

Fig -3: Object Detection – Project Flow

5. FACE RECOGNITION

Face Recognition refers to the detection and identification of the face automatically by computerized systems by looking at the face. With facial recognition, we not only recognize the person by drawing a box on his face, but we

also know how to give a specific name with OpenCV or Python. In this project, we have used OpenCV (Open-Source Computer Vision Library) library as it provides a Computer Vision library and image processing methods using machine learning. OpenCV was created with a heavy emphasis on real-time applications and was built for efficient calculation [9]. It is an essential module needed for face recognition using a camera.

Fig. 4 explains the phases involved in face recognition. The face from the dataset is matched to their corresponding user names. These names are converted to audio output after having text-to-speech conversion.

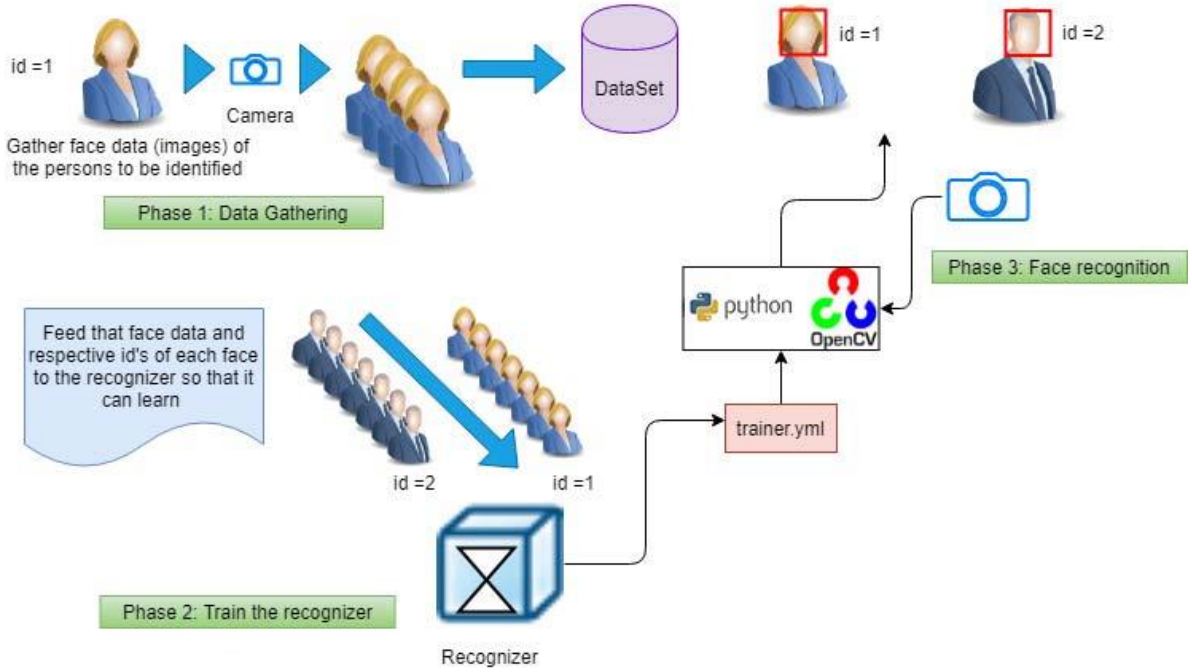

Fig -4: Phases of Face Recognition

Fig. 5 shows the flow of the face recognition model. The images of a person's face are added to the database in the training stage. Images are used for the feature extraction process and they are saved to the database. In the recognition stage, the video is acquired from the camera and it is converted to frames. Then the model uses the database and trained knowledge to recognize the face of the person. If the person's face is recognized, then his name is output in the audio format after undergoing conversion from text to speech.

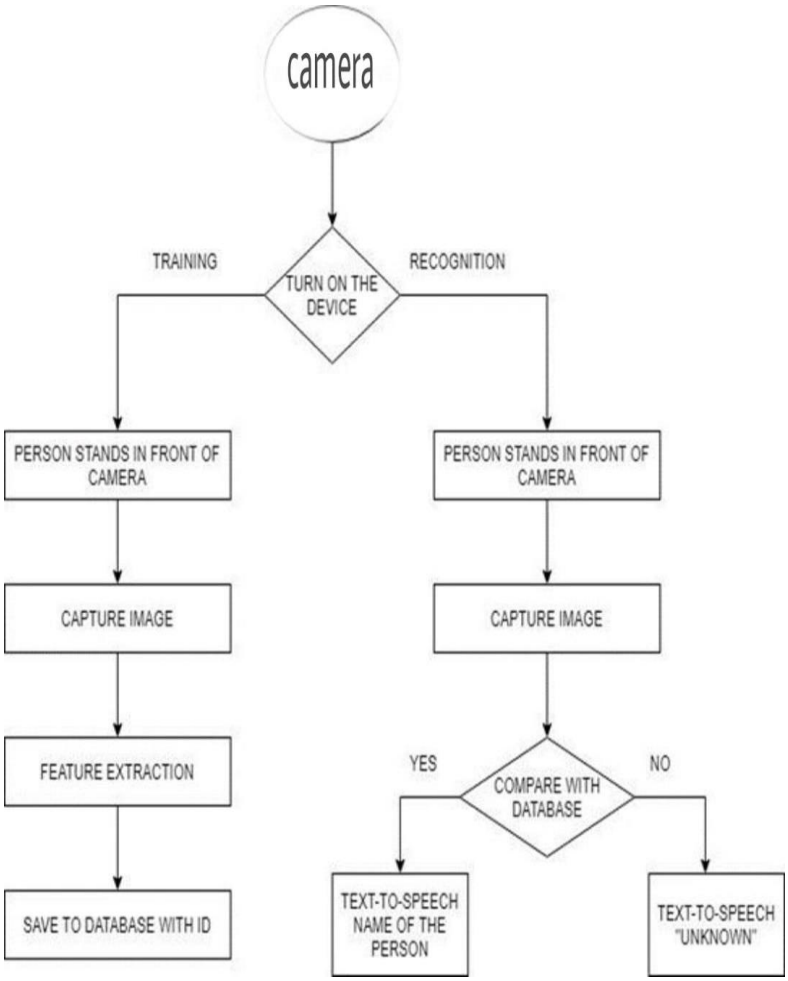

Fig -5: flow of face recognition model

5. GESTURE RECOGNITION

Computers can record and recognize human gestures as commands thanks to a perceptual computing user interface called gesture recognition. We used CNN approach to understanding hand gestures. A CNN (Convolution Neural Network) is a network architecture used for deep learning algorithms for image classification and recognition and tasks involving processing image data. In Fig. 6, the input is given to the CNN. It consists of layers. The convolutional layers are where the filters are applied to the original image. Various features and patterns in recognizing gestures require the usage of an input frame.

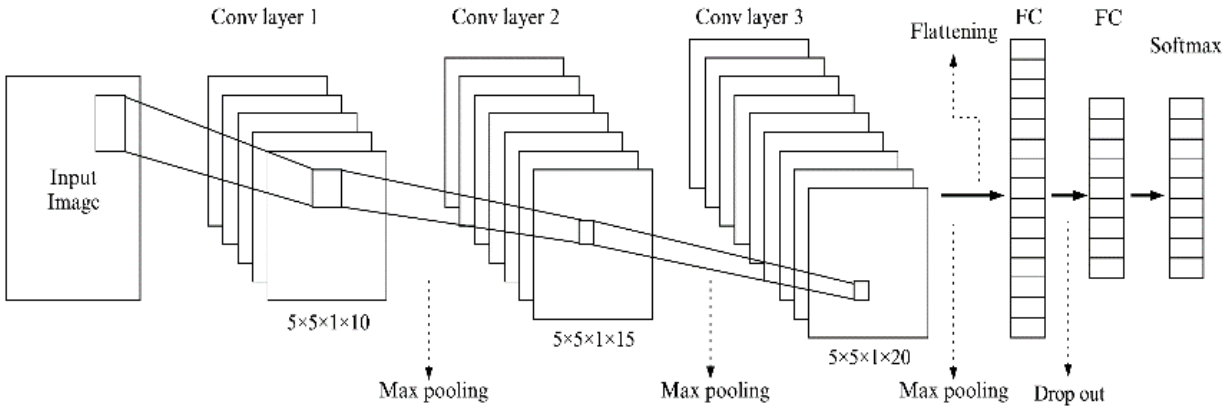

Fig -6: CNN architecture for Gesture recognition

The pooling layer is present after every convolution layer. The pooling layer reduces the number of the feature maps' parameters. This reduces the amount of computation in the network. The image classification happens at the Fully connected layers in the CNN based on the features extracted from previous layers. Fully connected layers are positioned before the classification output of a CNN and they are used to flatten the results before classification. The output from the fully connected layers is passed to a logistic function called the Softmax that produces a vector that depicts the probability distribution of a set of possible outcomes. The output class is the one with the highest probability. SoftMax function, a function that activates numbers (logits) to produce probabilities that add up to one [10].

5. CONCLUSIONS

This is a basic implementation of our project – Eyeris. The objective is to assist those who are blind to ease the difficulties faced by them. This project is capable of detecting and identifying objects that humans use daily. Our goal was to make it possible for those who are blind to do their daily tasks without the need for a guardian as much as possible.

Our project not only recognizes objects and people, however, it also enables the communication between deaf-dumb and blind persons. They can communicate through gestures or signs using our gesture recognition, where our model detects and identifies gestures performed by a deaf-dumb person and then outputs the gesture meaning to a blind person in the audio format. It also has a reading system where we can give images of pages as input and the system will read out the content in the image to the user. It detects the face of the person and outputs the person's name through audio if the individual has previously been saved on the database. After experimenting we found that our project is useful and can be applied in a real-time environment.

6. REFERENCES

[1]. P. A. K. S. M. B. Siddharth Pandey, "SMART VISION SYSTEM FOR BLIND", int. jour. eng. com. sci, vol. 3, no. 05, Dec. 2017.

[2]. R. Chellappa, C. Wilson, and S. Sirohey, "Human and machine recognition of faces: a survey," Proceedings of the IEEE, vol. 83, no. 5, pp. 705–741, 1995.

[3]. https://www.v7labs.com/blog/yolo-object-detection.

[4]. Global data on visual impairment," World Health Organization, 08-Dec-2017. [Online].

[5]. F. Henriques, H. Fernandes, H. Paredes, P. Martins, and J. Barroso, "A prototype for a blind navigation system," 2nd World Conference on Information Technology, Nov. 2011.

[6]. D. J. Mistry, V. V. Mukherjee, N. T. Panchal, and A. U. Gawde, "Oculus Vision for Blind," DJ ASCII-18, vol. 2, Jun-2018.

[7]. S. Ferreira, C. Thillou, B. Gosselin, "From Picture to Speech: An Innovative Application for Embedded Environment" in ProRISC 2003, Veldhoven, Netherland, 2003.

[8]. https://www.v7labs.com/blog/yolo-object-detection.

[9]. M. Naveenkumar and A. Vadivel, "OpenCV for Computer Vision Applications," Proceedings of National Conference on Big Data and Cloud Computing (NCBDC'15), Mar. 2015.

[10]. D. Jurafsky and J. H. Martin, "Logistic Regression," in Speech and Language Processing, 2019.