International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 10 ǁ October 2022 ǁ PP. 537-544

Virtual Smart Glass for Blind using Object Detection

Mohammed Shakkir P K, Basil Shaji, Muhammed Muhsin,Boby Martin Aprem, Dr.Abhiraj T K
Electrical and Electronics Engineering Ilahia College of Engineering and Technology Mulavoor, Kerala, India

Abstract-*One among the prevailing is that the vision disabilities are rapidly increasing. The disability of the person with vision difficulty is the inability to see and recognize people or objects. Therefore, the visually impaired person has to walk depending on a stick or dog or another person. The main motive of this project is to help the blind by modeling a device that makes them walk like normal people and to overcome their disability. This device, the virtual smart glass, assists them in their ways without the need of human help and helps them walk independently. The device is installed on the glass frame and these glasses help to figure out the surroundings. The object detection technology is given as input to the microprocessor, to find out if there is any obstacle in front of it. If there is an obstacle, the device will produce an audio form for alerting the user.*

Date of Submission: 10-10-2022 Date of acceptance: 22-10-2022

I. 1.Introduction

In recent days, many people suffer from different diseases or are disabled. NCBI says that 1.5% of the population in India are blind and 7.8% have vision difficulties. They require some help to make their life easier and comfortable. A new technology needed to be introduced to help them. Thus, the motive of Smart Glass is to aid them walk independently and freely. This system consists of wearable smart glass with object detection using neural network methodology to identify objects like person, rupees, QR, car, bus etc. Individual image processing description labels are used to detect the objects and are converted into voice commands. Therefore, smart glasses are a device which alters the wearer's vision, where one is physically located or where he/she looks.

There are several ways for differentiating the visual information which a wearer perceives. The human brain makes vision seem very much easier. The brain need not work hard to identify, apart from human identification, reading a sign, or recognizing a human's face. These are complicated to solve using a computer. They only seem easier just because our brains are good at understanding and recognizing the image. Image recognition and object detection is a software which helps in identification of objects, places, people, writing and actions in images. Computers use computer vision technologies in combination with a camera and artificial intelligence software to achieve image recognition. Software for image recognition requires deep machine learning. In convolutional neural networks, processor performance is best but requires massive amounts of power.

II. Proposed work

2.1 Image Input

The system detection is the task to identify images along with label boxes. System vision's deep learning algorithm possesses approx. 90 FPS, accurate SSD and Faster datasets. Database support for blobs is not universal. Hence there is no need to use server because the data is fixed, because the things which the smart glass captures and recognizes have particular feature and shape, for example, at home (tables, chairs, doors and so on) all these things have same feature in all places (in streets, cars, sidewalk, vehicle and so on). Therefore, the database is fixed on pattern recognition, so that it can be stored in a device which runs the application process for the eyeglasses.

2.2 Audio Input

The result of the detected objects and the distance it is in, are converted to text. Then using the prepared audio library for conversion of text to voices, the voice output is heard through the headphones.

2.3Tools Utilized

Stereo views pre-trained model along with the necessary packages has about 30 trained datasets useful to detect the object. Framework model used to create deep learning networks solves the object detection problem.

Some necessary Tools Utilized are:

1. Python 3.5
2. Caffe model framework
3. OpenCV
4. 4.NumPy1.14
5. Tensorflow lite

The pre-trained model of python OpenCV/Tensorflowlite , recognizes objects using a camera. It works based on real-time object detection with access to the minicamera in an efficient manner for each frame. The minicamera is pre-installed in the system. The live stream can be accessed with minicam output. The video frame detects the objects depending on the speed of the computer's CPU or GPU resources.
Anaconda command prompt can be used, but make sure that the anaconda commands prompt without changing the directories of the path.

2.4Object Recognition
The Object Recognition Workflow of the object detected model is created to avoid complexities in the system. The TensorFlow and open cv are models with dataset. The deep learning frameworks use libraries for real time computer neural networks in many real time applications like medical recognition, face detection, object detection.

2.5Speech Synthesis
The process of synthesizing text to voice comment is calibrated using Google's Text to Speech. API commonly used is the GTTS API which computes system speech synthesizer and is implemented in python library files. Thus the system converts the speech format into traditional speech text.

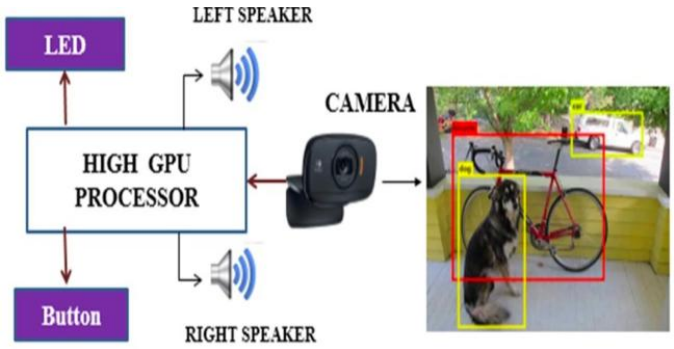

Fig. 2.1. Block Diagram of the Proposed Work

From the figure 2.1, Object detection images and videos are the input to the high GPU (raspberry pi 4) is a quad core processor and has roughly 50-60 percent better performance in 32-bit mode and 10x faster than the original single-core Raspberry Pi. Also, the camera placed in front of the camera slot, detects the images using the detection algorithm and glasses can capture images using the camera to recognize images and determine each object present in front by computer vision technique & pattern recognition system. The system can calculate the distance between the person and each specific object and convert the information to voice commands which is to be understood by the blind user via earphones and intimates them if they are very near to the object.

The proposed system is well trained and has a unique novelty in providing good accurate detection results in a simplified manner. The input of the virtual glass is taken from the output provided by the strip camera and provides subsequent feedback to the person through the earphones and thus it can help the blind while walking alone in an outdoor environment. This is a form of computing device, used and worn in the head.

When the person is wearing the glass moves from his other position, the camera adjusts such that it gets clear vision according to the position and orientation. Thus, the glasses are smart and unique compared to other computing devices of such kind. They also possess some unique features such as enabling new applications which are not easily adaptable compared to others.

III. Methodology

Initially the image is processed by several methodologies like detection, extraction, classification and finally the objects are being recognized. Binary images are processed in the form of pixels, then they are extracted and applied for real time binary image processing.

3.1Feature Detection

One of the methodologies employed in the system is called feature detection, where the computing abstractions of the specific set of image information occurs and detection by making decisions and recognition at every point of the image whether there is a feature of the image set or not. The resulting features are considered as the subsets of the image domain. Feature detection is mostly applied by four ways such as edges, corners (interest points), blobs (regions of interest point) and ridges.

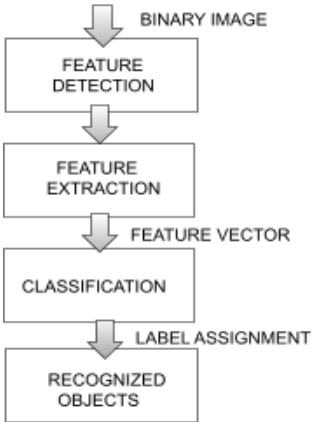

Fig.3.1.Object Recognition Methodology

3.2Edges

The boundary between two images is considered to be the edge. Generally, they should include junctions and are mostly arbitrary in shape. During the application, the edges are usually considered as a set of points within the particular image having a robust gradient magnitude. The Gradient points in the image are collectively chained together in order to make the image a more complete description of a foothold.

3.3Interest points

The corner or interest point is a methodology where the terms corners are referred to point-like features in a picture (fig 3.2). They have a two dimensional (2D) structure. The term "Corner" was first performed by the edge detection and then the edges were individually analyzed to find the rapid changes in direction or corners.
However, it was noticed that the corners were also detected on individual parts of the image which weren't corners in the sense.

Fig.3.2 Corner Interest points

3.4 Feature Extraction

The feature extraction is the method of extracting the useful and crucial information from the images. In this step, the image is transformed from input file to a group of specific features.

Features are nothing but distinctive properties of exclusive input patterns that help in differentiating between the categories of input by deriving new features from the existing initial features in order to scale back the cost and work of the feature measurement and to extend the efficiency of the image classifier with a higher order of classification accuracy. While extraction, if the features extracted are carefully chosen, the classification phase is reportedly proved to produce better results.

Several data analysis software packages exist in order to supply and support the feature extraction and also the dimension reduction. Therefore, for this process the numerical programming environments that are commonly used such as MATLAB and C language, supports a number of simpler feature extraction techniques. The fig.3.3 explains the feature extraction techniques.

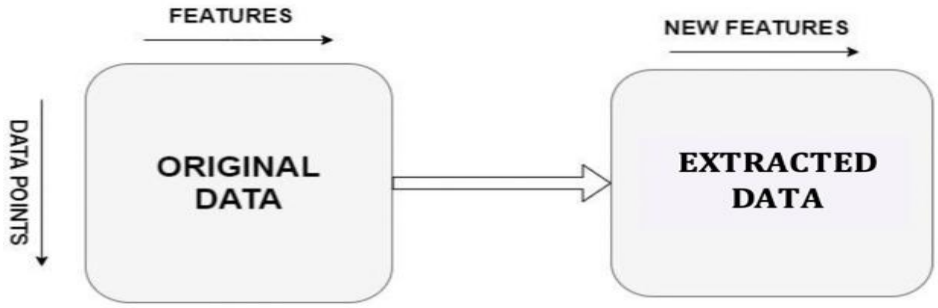

Figure 3.3. Feature extraction

3.4 Classification

Image classification is the methodology of recognizing the images using the feature vector which is the output of the feature extraction phase supported to the training set given to this program.Training samples can be easily created to represent the classes that have to be extracted, using the image classification toolbar.

This project uses the Speech Application Programming Interface or SAPI that allows the use of speech recognition and speech synthesis within the respective applications. There are several applications that use SAPI, such as MS Office, MS Agent and Microsoft Speech Server. The Speech API generally could be a free redistributable component. They can be transferred with a Windows application that uses speech technology. Many other versions of the speech recognition and synthesis engines are also available to freely redistribute.

3.6Text to Speech Conversion (TTS)

The text to speech conversion is built up though a nonfunctional COM framework. Various new steps are available for every new function demonstrating the use of XML tags to modify speech in sample. This step is to create the voice of assistance by simply declaring the instance and using CoCreateInstance where SAPI uses intelligent defaults. This generally requires a minimal amount of initialization where the voice can be used immediately.

In this, the defaults are located on top of things Panel in the speech properties which includes a variety of voices and several languages including the native ones such as English, Japanese, etc. Speech properties can be programmatically set to default values. The voices of the speech could also be modified using some sort of methods such as using XML commands on to the stream. In this situation, a relative rating will lower the pitch of the voice.

IV. Algorithm

YOLO: You only look once (YOLO) is a real-time object detection system. YOLOv3 is incredibly quick and precise. Besides, we can without much hectic tradeoff among speed and exactness basically by changing the size of the model, no retraining required! Earlier recognition frameworks repurpose classifiers or localizers to perform location. They apply the model to a picture at different areas and scales. High scoring areas of the picture are viewed as identifications. YOLO utilizes an entirely unexpected methodology. YOLO applies a single neural network to the entire image and divides the image into different parts. It also assigns different parts, unique weighted values by predicting the probability of different parts of the image. Because of this methodology it is highly efficient and faster as compared to some other object detection models. The model has

a few focal points over classifier-based frameworks. It takes a glance at the entire picture at test time so its forecasts are educated by worldwide settings in the picture. YOLOv3 makes use of the latest darknet features like 53 layers and it has undergone training with one of the most reliable datasets called ImageNet. The layers used are from an architecture Darknet-53 which is convolutional in nature. For detection, the aforementioned 53 layers were supplemented instead of the pre-existing 19 and this enhanced architecture was trained and instructed with PASCAL VOC. After so many additional layers the architecture maintains one of the best response times with the accuracy offered. It also is very helpful in analyzing live video feed because of its swift data unsampling and object detection techniques. One can notice that this version is the best enhancement in ML (Machine Learning) using neural networks. The previous version did not work well with the images of small pixels but the recent updates in v3 have made it very useful in analyzing satellite imaging even for defense departments of some countries.

SSD : The SSD approach is based on a feed-forward convolutional network that produces a fixed-size collection of bounding boxes and scores for the presence of object class instances in those boxes, followed by a non-maximum suppression step to produce the final detections.The SSD object detection composes of 2 parts:

- Extract feature maps, and
- Apply convolution filters to detect objects.

Multi-scale feature maps for detection :We add convolutional feature layers to the end of the truncated base network. These layers decrease in size progressively and allow predictions of detections at multiple scales. The convolutional model for predicting detections is different for each feature layer.

Convolutional predictors for detection :Each added feature layer (or optionally an existing feature layer from the base network) can produce a fixed set of detection predictions using a set of convolutional filters. These are indicated on top of the SSD network.For a feature layer of size m × n with p channels, the basic element for predicting parameters of a potential detection is a 3

× 3 × p small kernel that produces either a score for a category, or a shape offset relative to the default box coordinates. At each of the m × n locations where the kernel is applied, it produces an output value. The bounding box offset output values are measured relative to a default box position relative to each feature map location.

Default boxes and aspect ratios: We associate a set of default bounding boxes with each feature map cell, for multiple feature maps at the top of the network. The default boxes tile the feature map in a convolutional manner, so that the position of each box relative to its corresponding cell is fixed. At each feature map cell, we predict the offsets relative to the default box shapes in the cell, as well as the per-class scores that indicate the presence of a class instance in each of those boxes. Specifically, for each box out of k at a given location, we compute c class scores and the 4 offsets relative to the original default box shape. This results in a total of (c + 4)k filters that are applied around each location in the feature map, yielding (c + 4)kmn outputs for a m × n feature map. Our default boxes are similar to the anchor boxes used in Faster R-CNN however we apply them to several feature maps of different resolutions. Allowing different default box shapes in several feature maps let us efficiently discretize the space of possible output box shapes.

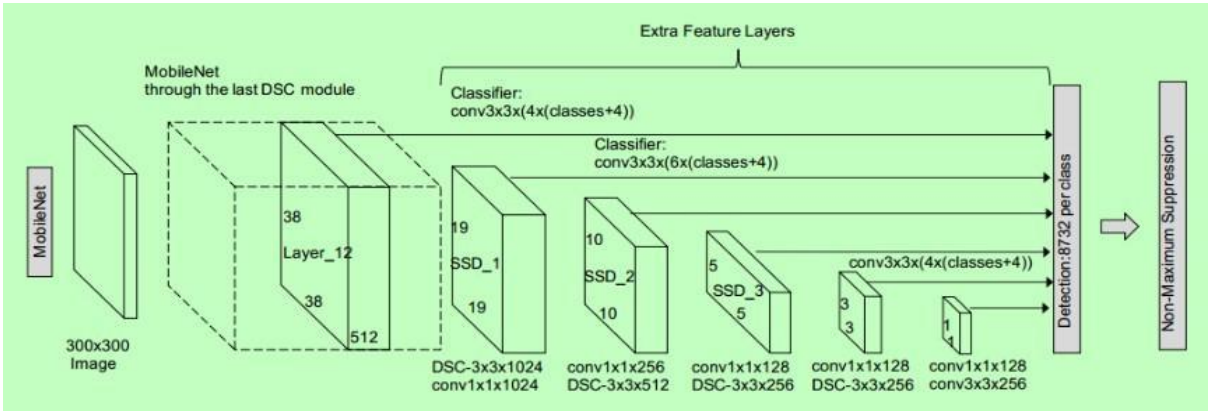

Figure4.1: SSD Single shot multibox detector

Mobilenet: MobileNet model is designed to be used in mobile applications, and it is TensorFlow's first mobile computer vision model.MobileNet uses depth wise separable convolutions. It significantly reduces the number of parameters when compared to the network with regular convolutions with the same depth in the nets. This results in lightweight deep neural networks.Adepthwise separable convolution is made from two operations:

- Depthwise convolution.
- Pointwise convolution.

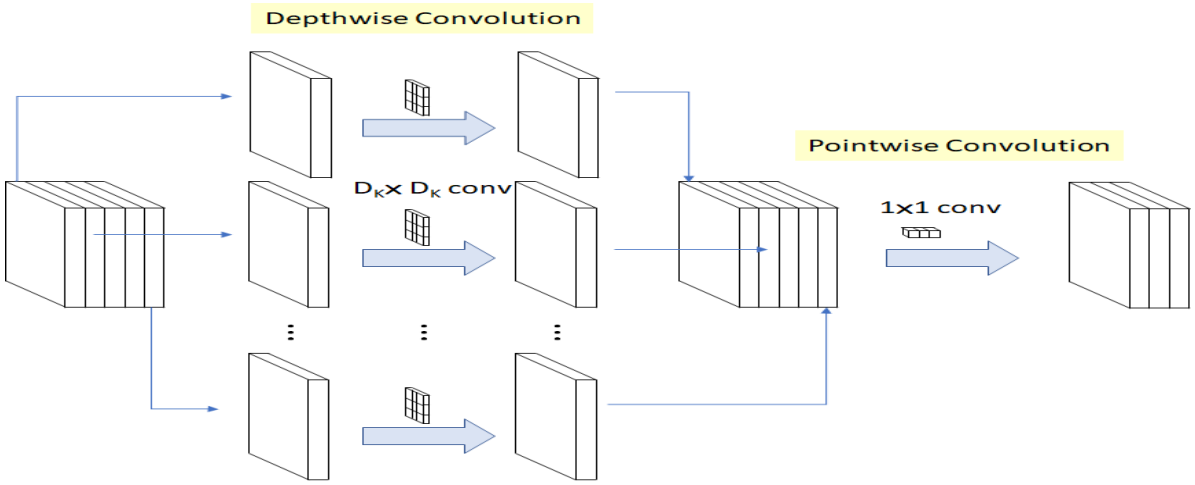

Figure 4.2: Depthwise separable convolution

1. Depthwise convolution is the channel-wise DK×DK spatial convolution. Suppose in the figure above, and we have five channels; then, we will have 5 DK×DK spatial convolutions.
2. Pointwise convolution is the 1×1 convolution to change the dimension

Depth-wise separate convolutions first apply a single filter on each input to filter the input data, followed by 1 × 1 convolutions which combine these filters into a set of output features. These depth-wise separable layers almost mimic the function of typical convolution layers but with much faster speed and with a slight difference (typical convolution filters and combines into output features both, but in depthwise separable convolution this is divided into two layers, one separate layer for filtering and one separate layer for combining). This minimizes the model size and reduces computational power demands. All layers are followed by a batchnorm and ReLU nonlinearity except the final fully connected layer which feeds into a softmax layer for classification having no nonlinearity. MobileNet has 28 layers without Counting depthwise and pointwise convolutions.

Mobilenet SSD: Generally, SSD uses an auxiliary network for feature extraction. This is also called as base network.There are some practical limitations while deploying and running complex and high power consuming neural networks in real-time applications on cut-rate technology. Since, SSD is independent of its base network, MobileNet was used as the base network of SSD to tackle this problem.This is known as MobileNet SSD.

When MobileNet V1 is used along with SSD, the last few layers such as the FC, Maxpool and Softmax are omitted. So, the outputs from the final convolution layer in the MobileNet is used, along with convolution a few more times to obtain a stack of feature maps.These are then used as inputs for its detection heads. Its architecture can be modified as per required.The table below gives one of its architecture in detail.

Google text to Speech API: A Text-to-speech API is an technique which makes use of the natural language processing techniques to convert the text samples into speech signals,by analyzing and processing the text and at that point utilizing Digital Signal Processing (DSP) innovation to change over the processed content into synthesized speech representation of the content. This API will be useful if the person is blind, just projecting the text won't solve the problem. For him, special algorithms, which convert the processed text to speech again, are required to be outputted through the speaker. This API improves customer interactions with intelligent, lifelike responses,engages users with voice user interface in your devices and applications and personalizes communication based on preference of voice and language.

V. Hardware Description

RaspberryPi-Raspberry Pi is a credit card sized microcontroller which acts as a minicomputer that connects all peripherals used by computers like keyboard, mouse, TV or monitor etc. It is a low costmicrocontroller that helps people of all ages to explore computing. Raspberry pi has an SD card slot for

loading operating, Ethernet port for LAN cable connection, power supply, 4 USB ports for connecting I/O devices, memory, audio/video outputs and camera interface which can be connected to a Raspberry Pi camera. General purpose input output (GPIO) pins are used to connect the controller with Raspberry Pi. The GPIO pins can be used for both input and output. In our case, we use them for input; the buttons of the controller are connected to these pins through which the input is taken. The earphones and Raspberry Pi camera are connected to the board. The camera is connected to the camera slot and earphones to the jack slot. Figure shows the Raspberry Pi 4 microcontroller in which the parts of the microcontroller are labeled. The operating system (OS) used with Raspberry Pi in our project is Raspbian OS which is a Debian-based free operating system. Raspberry Pi camera is used for taking the real time input for detecting obstacles and recognizing faces. The Raspberry Pi camera is a second generation module which is portable and lightweight that supports Raspberry Pi and has a fixed focus lens. The camera is normally used for image processing and machine learning projects. Fig 5 shows the Raspberry Pi camera which can be attached to Raspberry Pi through the camera interface on the microcontroller.

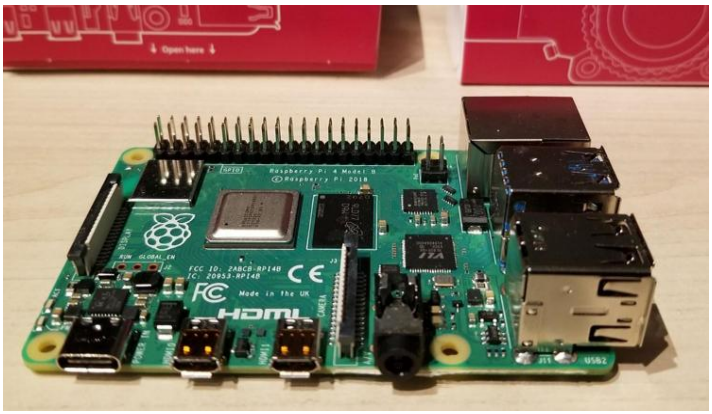

Figure.5.1. Raspberry pi 4

Camera Module: The Pi camera module is a portable lightweight camera that supports Raspberry Pi. It communicates with Pi using the MIPI camera serial interface protocol. It is normally used in image processing, machine learning or in surveillance projects. It is commonly used in surveillance drones since the payload of the camera is very less. Apart from these modules Pi can also use normal USB webcams that are used along with computers.

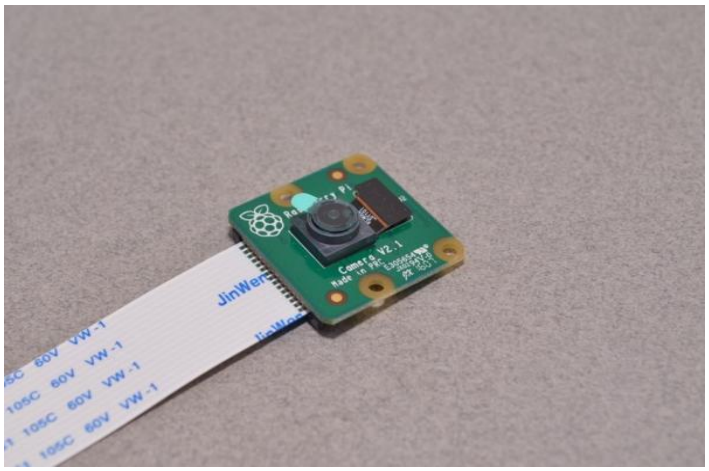

Figure 5.2. Pi Cam

VI. Future scope

In the future for blind people and people who have vision difficulties by adding new techniques. For instance, direction and warning messages to prevent expected accidents, messages to tell the user about the battery level, video detection to provide a full healthy life for people with vision difficulties, develop mobile application to control "Smart Glasses", use 270 camera to have more wider view angle., provide the glasses with GPS notification and develop the glasses' design to have little, small and light components so the user can wear it easily.

We could implement the same concept using a smartphone, so that we can avoid the use of GSM,

GPS and Raspberry Pi zero modules. Also, we will be implementing the voice command in an advanced way using any of the available platforms like, i.e. google assist, Siri, Cortana, Bixby, Alexa. Also, the overall design and casing of the smart glass will be improved to achieve a very compact form.

This project could interface the details of unknown persons from the government to our projects to know the names of unknown people in front of blind persons. This helps blind people a lot in identifying the persons in front of him. We expect further improvements in the future as we develop new feature types including color, distance and other features.

VII. Conclusion

The main objective of this paper is explained briefly, which is the need for a voice assistant system for the blind people all over the globe. The "Virtual Smart Glass for Blind" is a practically feasible device which can be easily worn and used by any person.The detected images of the objects in front of the camera and these detected images are then converted to audio and fed as audio. Thus, the smart glass will be a major help for the blind society in guiding them.

References

[1]. Third Eye for Blind, Sandhya B R, S Sahana. Information Science & Engineering, Sri Krishna Institute of Technology, Bangalore, India , DOI:https://doi.org/10.5281/zenodo.4592708

[2]. Smart Glass Using IoT and Machine Learning Technologies to aid the blind, dumb and deaf, S Salvi, S Pahar, Y Kadale - Journal of Physics: Conference Series, Volume 1804, InternationalConference of Modern Applications on Information and Communication Technology(ICMAICT) 22-23 October 2020, University of Babylon, Babylon-Hilla City, Iraq, 2021 - iopscience.iop.org

[3]. Smart Glass System Using Deep Learning for the Blind and Visually Impaired, MMukhiddinov, J Cho - Electronics, 2021

[4]. MayureshBanne, Rahul Vhatkar, RuchitaTatkare(2020), Object Detection and Translation for Blind People Using Deep Learning, International Research Journal of Engineering and Technology (IRJET), Volume: 07 Issue: 03.

[5]. M . M u r a l i , S h r e y a S h a r m a , N e e l N a g a n s u r e (2 0 2 0) , Reader and Object Detector for Blind, International Conference on Communication and Signal Processing, July 28 - 30, 2020, India.

[6]. Mohammad Marufur Rahman Milon Islam Shishir Ahmmed , Saeed Anwar Khan(2020), Obstacle and Fall Detection to Guide the Visually Impaired People with Real Time Monitoring, SN Computer Science (2020) 1:219

[7]. V.Balaji, S. KanagaSubaRaja , C.J. Raman , S.Priyadarshini , S.Priyanka5 , S.P.Salaikamalathai,(2020), real time object detector for visual impaired using open cv, European Journal of Molecular & Clinical Medicine ISSN 2515-8260 Volume 7, Issue 4, 2020.

[8]. Lijun Yu, Weijie Sun, HuiWang,Qiang Wang and Chaoda Liu(2018), The Design of Single Moving Object Detection and Recognition System Based on OpenCV, Proceedings of 2018 IEEE International Conference on Mechatronics and Automation August 5 - 8, Changchun, China.