PROJECT : ABILIS

Project done under Artificial Intelligence and Electronics Society(ArIES), IIT Roorkee

Mentees:-

Srushti Borate (20115023)
Harsh Chauhan (20115037)
Nitya Tyagi (20311017)
Vidhi Patidar (20311025)
Ganesh Kalanidhi (20110012)
N.Arutkeerthi (20311016)

Mentors:-

Manav Saraf
Nitika Gupta

Overview

This is a vision enhancer based module specifically designed for the Visually challenged people. The system is designed in such a way in which the blind person can take the help of an application which sends Real Time Frames to the system. It works on real-time object detection using SSD_MOBILENET algorithm and TENSORFLOW APIs. In image classification, an entire image is classified. Object detection extends image classification by detecting the location of individual objects present in an image. And with the help of semantic segmentation, we extract the walkable paths and hence it helps in navigating the blind people.

Importance

The ability to navigate from place to place is an integral part of daily life. We understand the need for assistive devices for navigation and orientation for visually challenged people and hence we decided to take this on us to do our bits to help in

making their life a bit easy. And we believe that technology truly has the ability to empower everyone without any bias. We came up with this idea to design products (precisely glasses) specifically for visually impaired people. These designs will make a huge difference in their lives. Blind individuals or people with low vision can perform everyday tasks independently with the help of such innovations.

Features

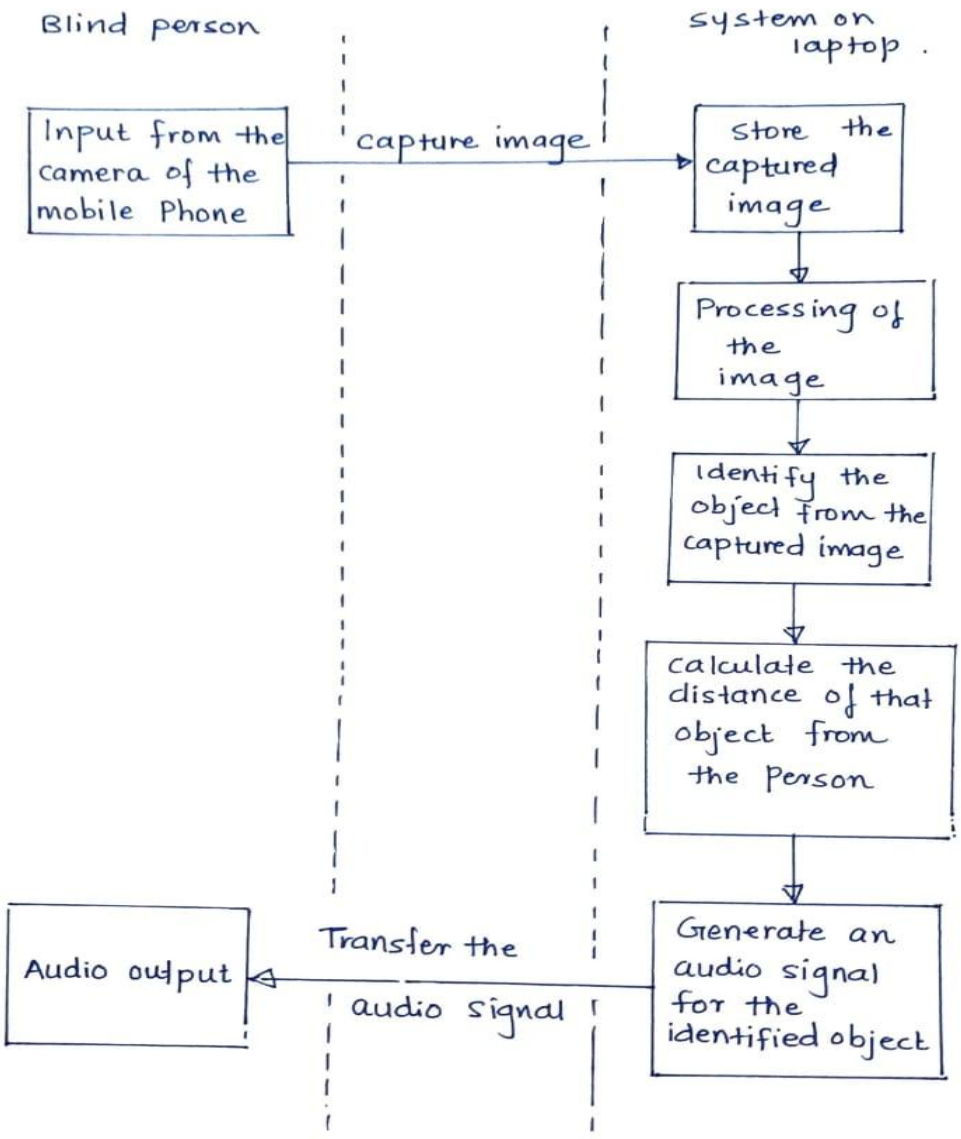
- Easy to integrate
- Voice Activated
- Real Time Interface
- Can be Interacted with Android Mobile Webcam as well

Action Plan

We are taking input(video feed) in real time and it undergoes two processes, first is object detection and second is semantic segmentation and generating a voice feedback as output.

(I) Object detection

Technical setup


Blind person
system on laptop.
Input from the camera of the mobile Phone
capture image
Store the captured image
Processing of the image
Identify the object from the captured image
calculate the distance of that object from the Person
Generate an audio signal for the identified object
Transfer the audio signal
Audio output

1. Methodology

(A) Depth estimation

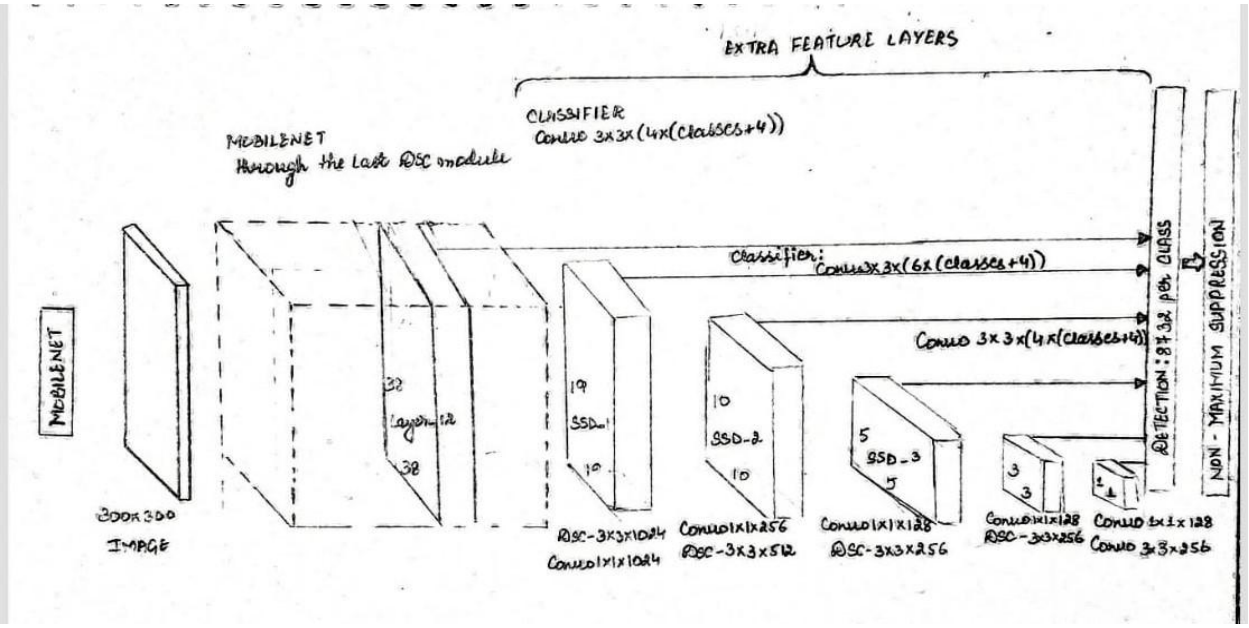
It is used to calculate distance between two real time objects. We need to find out how far the obstacles and people are located in any real time situation. After the object is detected, a rectangular box is generated around the object. If that object occupies most of the frame then with respect to some constraints, the approximate distance of the object from the particular person is calculated.

(B) SSD_MobileNet

SSD has two components: SSD head and a backbone model. Backbone model basically is a trained image classification network as a feature extractor. Like ResNet, this is typically a network trained on ImageNet from which the final fully connected classification layer has been removed.

The SSD head is just one or more convolutional layers added to this backbone and the outputs are interpreted as the bounding boxes and classes of objects in the spatial location of the final layers activations.We are hence left with a deep neural network which is able to extract semantic meaning from the input image while preserving the spatial structure of the image albeit at a lower resolution.For an input image ,the backbone results in a 256 7x7 feature maps in ResNet34 . SSD divides the image using a grid and has each grid cell be responsible for detecting objects in that region of the image. Detecting objects basically means predicting the class and location of an object within that region.

The SSD_MobileNet model is based on depth wise separable convolutions which are a form of factorized convolutions. These factorize a standard convolution into a depthwise convolution and a 1 × 1 convolution called a pointwise convolution. For it, the depthwise convolution applies a single filter to each input channel. The pointwise convolution then applies a 1 × 1 convolution to combine the outputs of the depthwise convolution. The depthwise separable convolution splits this into two layers – a separate layer for filtering and a separate layer for combining. This factorization has the effect of drastically reducing computation and model size.



(C) Voice Module/Feedback

Voice generation module warns or signs the blind people by generating audio commands which are essentially understood by them.This is achieved with the help of certain libraries like pytorch, pyttsx3, pytesseract and engine. Pyttsx3 is a text-to-speech conversion library in python.Sometimes there is a need to identify the hidden text in the image, for this purpose Python-tesseract is used.

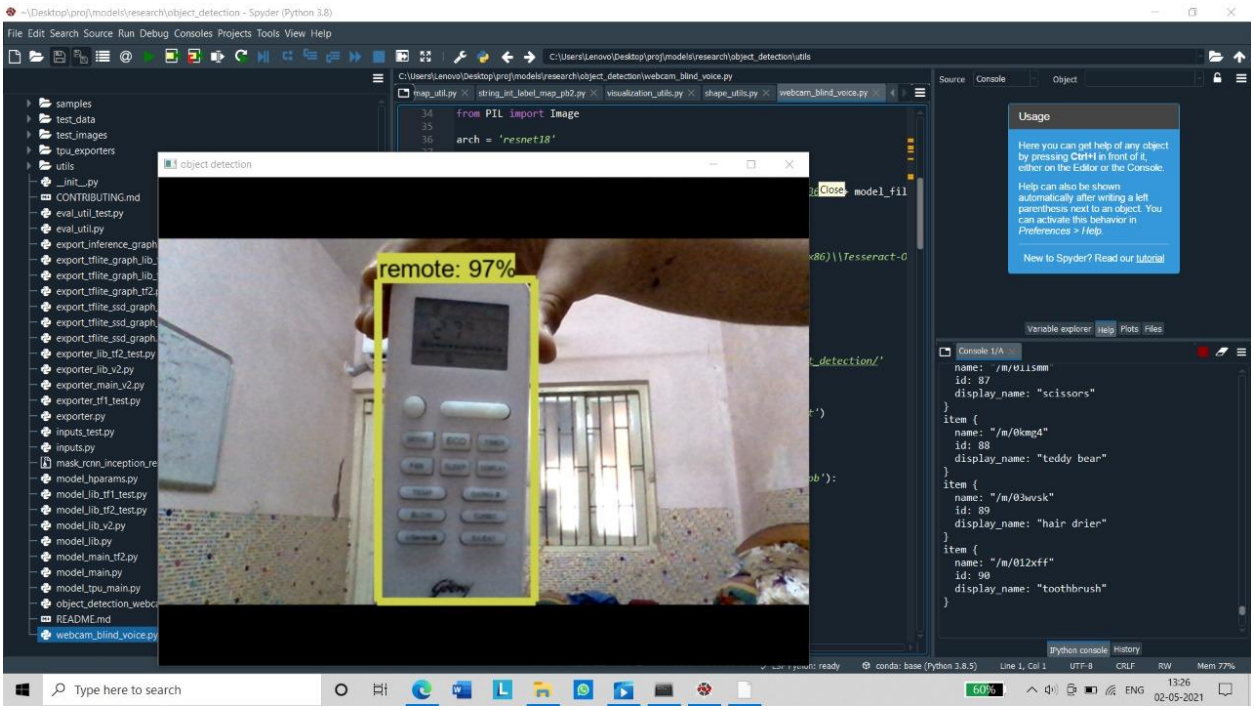
Audio commands are generated as output. If the object is too close then it states "**Warning: The object (class of object) is very close to you. Stay alert!**". Else if the object is at a safer distance then a voice is generated which says that "**The object is at a safer distance**". This is achieved with the help of certain libraries like pytorch, pyttsx3, pytesseract and engine.io **.**

(D) Reference source Code

https://github.com/beingaryan/Blind-Assistance-Object-Detection-and-Navigation

OUTPUT

Below are the objects on which it was tested and it gave the following result which were analysed further with the help of matplotlib libraries.

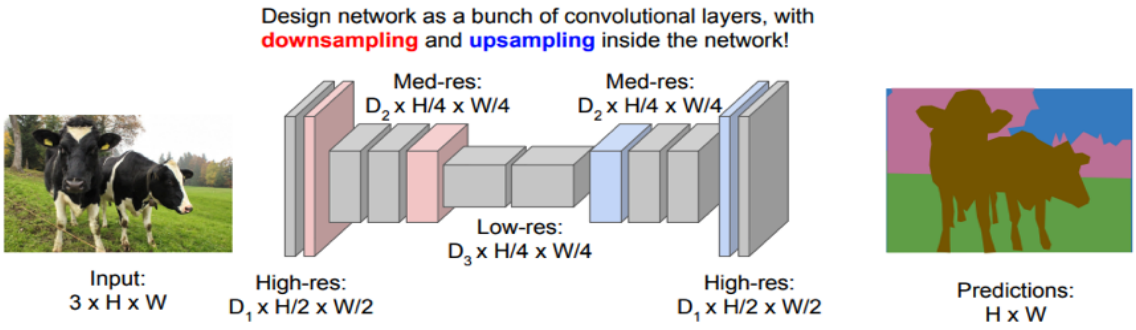


Consoles View

It calculates the distance and shows it’s output on consoles and the blind person can hear these distance coordinate values along with object detection based warning in a voice feedback.

```
0.0
Warning -Person very close to the frame
0.0
Warning -Person very close to the frame
0.0
Warning -Person very close to the frame
0.2
0.3
0.3
0.3
0.3
0.3
Warning -Person very close to the frame
```

(II) Semantic Segmentation:



Solution: Make network deep and *work at a lower spatial resolution* for many of the layers.

So far, our aim is to recognize the walkway on which the blind person has to move. We took images of the path ,reduced the picture size and cut its center 64x64 . The direction network (it is a 3-class classifier) commands to drive left or right or straight according to the path segmented.

Every image is made up of a group of pixel values. Image Segmentation is the task of classifying an image at the pixel level. A machine is able to analyse an image more effectively by dividing it into different segments according to the classes assigned to each of the pixel values present in the image.
Objects classified with the same pixel values are segmented with the same colormaps.
With help of semantic segmentation, we extract the walkable paths and hence it helps in navigating the blind people.

Reference source code taken from:

https://github.com/selectstarofficial/segmentation-selectstar

Aim: our aim is to recognize the walkway on which the blind person has to move. We took images of the path ,reduced the picture size and cut its center 64x64 . The direction network (it is a 3-class classifier) commands to drive left or right or straight according to the path segmented.

Model used Deep Lab V3+

Fully Convolutional Neural Networks are often used by semantic segmentation. One challenge of using FCN's for segmentation tasks is that the input feature maps become smaller and smaller while traversing through the convolutional and pooling layers leading to loss of information about the images leading to predictions of low resolution. Hence The DeepLab model addresses this challenge by using Atrous convolutions and Atrous Spatial Pyramid Pooling (ASPP) modules.

Tools used : Anaconda prompt and Spyder IDE

The code defines seven classes for segmentation but we are concerned with the class: sidewalks. For detecting the sidewalks out of the input we apply a mask over the sidewalk region.(Here a green segment is used to differentiate the walkable path). Once this mask (segmap) is created it is interpolated with the input path the overall result is given as the output.

Now how do we find the path direction??

What we do is we extract the lowermost row of the segmap array which contains all the information about the pixel values of the segment. The segmap array is a tensor which contains three red, green, blue channels. The Lowest row for all the three

channels is the added as f_row in the code. This f_row is divided from the center assuming our camera focus to be at that point. Then the sum of the pixels in f_row to the left and right are compared. Evidently if the sum of pixels on right is greater than that on left the path is heading towards right and vice versa.



Original input path segmap Final output

Testing

Third Party App provides ease and freedom in the field of app development. It brings efficiency and also helps in fast delivery of the output. Third Party App allows you to divide your work in parts and helps you to focus on the core part of the app or any system. This strategy helps in the development of good and quality software. We can pass on the Features of the Third Party App to the system.

1) At first, we are capturing real time images from the rear camera[4] of the mobile handset of blind people and a connection is established between mobile phone and system in laptop and then those images are sent from the mobile phone to laptop.

2)This connection is done by a Third party app which is installed in the mobile phone of the person. All the real time images which get captured by the rear camera of the mobile phone are first transferred to the Third party app in the mobile phone and then those images are sent to the laptop where they are processed for some further conclusions.

3) The system in the laptop will test it using its APIs and SSD_MobileNet ALGORITHM and it detects the confidence accuracy of the image which it is testing. We reached upto 99% accuracy for certain classes like books, cups, and remote.

4) After testing the images, we are generating an output on the laptop based system and its prediction is being translated into voice with voice modules and sent to the blind person with the help of wireless audio support tools.