An Effective Navigation System Combining both Object Detection and Obstacle Detection Based on Depth Information for the Visually Impaired

Authors

Md. Nishat Raihan
Student ID: 144413

Hossain Mohammad Seym
Student ID: 144418

Supervisor

Rafsanjany Kushol
Lecturer
Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT)

Department of Computer Science & Engineering (CSE)
Islamic University of Technology (IUT)
Organization of the Islamic Cooperation (OIC)
Gazipur, Bangladesh

October 17, 2018

Declaration of Candidates

This is to certify that the work presented in this thesis is the outcome of the analysis and investigation carried out by the candidates under the supervision of Rafsanjany Kushol in the Department of Computer Science and Engineering (CSE), IUT, Dhaka, Bangladesh. It is also declared that neither of this thesis nor any part of this thesis has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

- - - - - - - - - - - - - - -
(Signature of the Candidate)

Md. Nishat Raihan
Student ID: 144413
Academic Year: 2017-18
October, 2018

- - - - - - - - - - - - - - -
(Signature of the Candidate)

Hossain Mohammad Seym
Student ID: 144418
Academic Year: 2017-18
October, 2018

- - - - - - - - - - - - - - -
(Signature of the Supervisor)

Rafsanjany Kushol
Senior Lecturer
Department of Computer Science & Engineering (CSE)
Islamic University of Technology (IUT)
October, 2018

Abstract

Object detection remains one of the most researched areas in the field of Digital Image Processing. With the introduction of Convolutional Neural Network (CNN), there has been a revolution in the detection approaches. Although the detection algorithms have come a long way, detecting objects for the blind or visually impaired people (BVI) is a completely different scenario. Rather than the detection of objects, for the visually challenged people this task is primarily focused on obstacle detection. Based on this concept, several approaches have been made to design smart canes that can be used as a helpful walking tool. More robust approaches include real time imaging through camera devices and processing the images to detect objects or obstacles. It remains a challenge to ensure both sufficient performance and cost efficiency at the same time. In many cases, the design architecture is not convenient enough for the visually handicapped persons. Also very few attempts were made to combine depth information with an object detection method in real time.

In this paper, we propose a completely new system framework that performs detection for the visually challenged people. We use a depth sense camera and a portable computing device to analyze the depth data and combine with the detection method to detect objects and also obstacles in real time along with its relative position and also the distance from the user. We perform the object detection using YOLO (You Only Look Once) algorithm which is comparatively faster than almost any recent object detection algorithm. Even if an object is not detected by YOLO due to lack of light or any other cause, the depth information will allow us the detection of obstacle and also the position and distance can still be calculated. Finally the total information gathered in real time will be narrated with convenience to the subject.

Contents

1 Introduction

1.1 Overview

According to the the World Health Organization (WHO), as of 2017, approximately 253 million people in the world live with some sort of vision impairment. Among them, 36 million are blind and 217 million have moderate to severe vision impairment. It is a challenge for the Blind or Visually Impaired (BVI) people to perform several of their daily activities properly. Moving outdoors without assistance and in unknown environments poses a difficult situation for them. Several sensor based and computer vision based systems were developed over the years to aid them while moving on their own. Also while moving around, rather than only providing them the assistance to find a pathway, avoiding the obstacles, it is also very helpful and convenient approach to provide them the scene analysis information.

For decades, white cane remains the symbol for the blind or visually impaired people. With the advancement of technologies, canes have been designed to be smarter and be of more assistance than before. Canes are normally designed to grasp the idea of the obstacle through physical contact with the object. To ensure detection without contacting physically, depth sensors are used with the canes that are able to provide an abstract idea of distance to the subject.

Figure 1: Example of a BVI people with Visual Aid.

Computer Vision based approaches have also come to the aid of the BVI people in recent years. Computer Vision based methods usually focus on either Purposeful Navigation or Object Identification. The system frameworks based on purposeful navigation mainly focus on the detection of the obstacles. The challenge is to keep detecting the obstacles while providing sufficient feedback so avoiding obstacles becomes possible and the subject can reach the destination without any collision in the way. On the other hand, the system frameworks based on object identification primarily focuses on detection of the objects and provide the feedback to the subject in real time. In both scenarios, there are certain hardware requirements that can vary from one system design to another. But almost in all cases, either one to multiple depth sensors or depth sense camera devices are needed. A portable computational device might also be needed for real time processing and power supply. These requirements are necessary for methodologies that perform real time object detection.

1.2 Problem Statement

As object detection algorithm is not developed keeping in mind that it will be used for the visually impaired. So, not any algorithm can directly be implemented for a visually impaired person. Again, all the information in an image is not of concern for a visually impaired person, as he should not be interested in any object that is 100ft away from him. Rather objects within a certain range will be the prime concern and detecting them with accuracy is the real challenge. After the detection is done, the class labels of the detected objects along with their positions with respect to the user should be presented conveniently to the user.

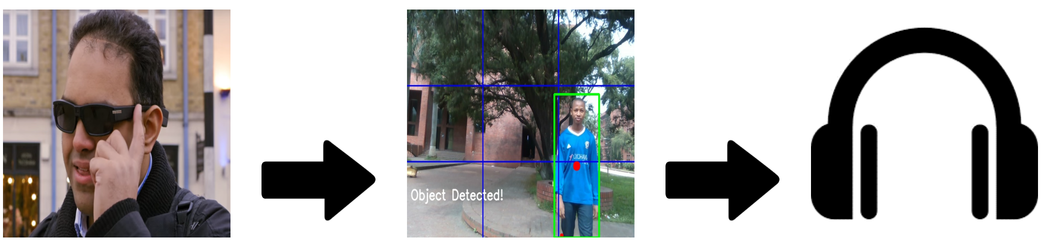

Figure 2: The complete detection and narration scenario.

1.3 Research Challanges

It is a challenge to implement object detection algorithm in a system that is intended for the BVI people. As object detection algorithms are not developed keeping in mind that it will be used for the visually impaired. To implement an object detection algorithm for the BVI people, it has to be fast enough to perform in ral time. Also, almost all detection algorithms require high end computational devices to perform efficiently, which poses a huge challenge to the portability of the system. Again, the detection method has to be compatible with the depth information provided by the depth sensors or depth sense cameras to perform altogether. So, not any detection algorithm can directly be implemented for the BVI people. Again, all the information in an image is not of concern for them, detecting objects that are comparatively away is a lesser concern for them. Rather objects within a certain range will be the prime concern and detecting them with accuracy is the real challenge. After the detection is done, the feedback to the subject should be convenient enough. The feedback can be in the form of an audio instruction including the names or tags of the detected objects along with their relative position and distance relative to the user. Major challenges are as follows:

1. **Ensuring speed on the portable device:** Most object detection algorithms require high end computational requirements to perform effectively. On the other hand, NAVI systems has to be portable. Ensuring portability for heavy configuration computing devices is a tough task.

2. **Performing in Real Time:** NAVI systems has to be real time. Although recent object detection algorithms perform fast enough, detecting obstacles can be done in a far shorter amount of time.

3. **When Detection Algorithm does not Detect Anything:**In case of obstacle detection, there is not much necessity of classification. Obstacles are detected based on distance information. But, while object detection, classification is a must to give the object a class name. If the detection algorithm fails to classify any object, it will not be able to give any feedback to the user.

4. **Distance for Overlapping Objects:** As depth is important for calculating the distance between the object and the user, it is easy when the objects are separated in the image. But when two objects overlaps, it gets difficult to detect their distances separately and uniquely.

5. **Power on the Portable Device:** Charging and recharging the device is also a challenge, as the continuous detection is quite power consuming. Providing power supply for the while of it's use is a tough scenario.

6. **Wearing the Deviec:** As the user will be visually impaired, it is expected that he/she will wear the device for convenience. The way we are proposing this, the device can not be attached to the cane, it should be done otherwise but also must ensure the comfort and ease of use for the user.

1.4 Thesis Objectives

Already we have discussed the research challenges for this work. There has not already been any work directly on implementing the fast and heavy object detection techniques for the visually impaired ones. As, we do not have to focus on the background objects and we are using a depth camera to gain the depth information, we have enough scopes to modifying the detection algorithms to work faster. Combining this with a convenient narrator and it can be very much efficient for the visually impaired persons.

- Our main objective of this thesis is to develop a methodology for detecting objects in digital images which can ensure real time detection in spite of:
  - noise
  - different scaled
  - overlapping objects
- In case, the object detection is not possible, we will try to perform obstacle detection instead.
- Design a complete detection system that can also calculate the depth and position relative to the observer.
- As the subject is expected to be BVI, to narrate the information with convenience so that it is easily understandable.

1.5 Thesis Contributions

In our work, we propose a complete system framework that is designed to work in real time with minimum need of user command while in use. We combine both object detection and obstacle detection based on depth information. For depth data Intel RealSense Depth Camera D435 [1] for gaining the depth information of a regular RGB image. The D435 is light-weight and easily movable on the outside. It ranges from 0.1 to 10+ meters and also works in sun the light. So, it meets our requirements without any complications. As the total system framework is expected to build such a system that can easily be carried outside without any discomfort, we used an Intel UP Board [2] for the processing power which will work as the portable light-weight computer, attached to the D435 camera to complete the total system. Although the Up Board ensures portability more, the D435 can also easily be used with any regular configuration laptop computer too.

The contribution of our work can be summarized as follows:

1. We proposed a completely new system architecture to aid the BVI people with better portability,performance,convenience and robustness.

2. We combined both object and obstacle detection to perform as one complete system.

3. We combined the detection methods with depth data to calculate the position and distance of the object in real time compared to the subject.

4. We provide real time feedback with necessary information with convenience and continues without any user interpretation.

5. The whole system is designed to perform without the need of any server.

1.6 Organization of the Thesis

The rest of the thesis will be organized as follows: in Chapter 2 we present the literature review of existing methods and their performance as well as limitations for the detection process. In Chapter 3, we propose our detection method for copy-move forgery. There we discuss about the overall idea of our proposed method and step by step implementation process. In Chapter 4, experimental set up, experimental result and performance analysis of our proposed method with various challenges are shown. Besides with other methods a comparative analysis is also shown. Finally, in Chapter 5, we conclude our thesis contributions and shows the future scopes for further developing the proposed method.

2 Literature Review

2.1 Navigation Methodologies for the BVI

Autonomous navigation is of extreme importance for those who suffer from visual impairment problems. Without a good autonomy, visually impaired people depend on other factors or other people to perform typical daily activities. Within this context, a system that can provide robust and accurate obstacle detection in urban environments, like city or indoors, is much more than desirable. Nevertheless, a major limitation of these systems is the usual distrust of visually impaired community towards the new technologies. As a consequence, this work proposes a user study with visually impaired people in order to obtain relevant feedback information about the system. In addition, the proposed obstacle detection algorithm can be easily integrated into more advanced vision-based localization systems for the visually impaired [3] [4].

Nowadays, most of the commercial solutions for visually impaired localization and navigation assistance are based on the Global Positioning System (GPS). However, these solutions are not suitable for the visually impaired community mainly due to low accuracy, signal loss and the impossibility to work on indoor environments. Moreover, GPS cannot provide local information about the obstacles in front of or in the near surroundings of the person. Furthermore, other commercial products available in the market present limited functionalities, have low scientific value and are not widely accepted by the users [5].

Computer vision-based approaches offer substantial advantages with respect to those systems and constitute a promising alternative to address these problems. By means of visual Simultaneous Localization and Mapping (SLAM) techniques[3] [6], it is possible to build an incremental map of the environment, providing at the same time the location and spatial orientation of the user within the environment. In addition, compared with other sensory modalities, computer vision can also provide a very rich and valuable perception information of the environment such as obstacle detection [7] or 3D scene understanding.

2.2 Related Works

2.2.1 Object Detection Algorithms

Object Detection and Recognition has recently become one of the most exciting fields in computer vision and AI. The ability of immediately recognizing all the objects in a scene seems to be no longer a secret of evolution. With the development of Convolutional Neural Network architectures, backed by big training data and advanced computing technology, a computer now can surpass human performance in object recognition task under some specific settings, such as face recognition.

The following detection algorithms are most notable mentions in this field in recent years:

2.2.1.1 Alex Net[8]:

- CNN is used for the first time in the field of object detection.
- CNN is designed consisting of five convolution layers and three fully connected layers with some customized modifications.
- To reduce training time, they used max(0,x) instead of traditional ReLUs like sigmoid function or tan h to reduce error rate they used overlapped pooling with stride = 2 and receptive field of 3*3.
- For the Overfitting issue, Data Augmentation and Dropout Method has been applied.

- **Limitations:** Takes too much time to train, the more conv layers are added the more training time it needs, accuracy not high enough, can not localize.

2.2.1.2 Inception Network[9]:

- Rather than using any fixed size kernel, Inception used all 1*1,3*3,5*5 kernels parallelly along with 3*3 max pooling. All of them are concatenated later on.
- To reduce the number of operations due to larger kernels like 3*3 or 5*5, first 1*1 kernels were used following 3*3 or 5*5. For max pooling, 1*1 kernels are used afterwards.
- Multiple Inception modules are placed altogether to form an Inception Network.
- In the network, there are side branches to check, if the network is already able to predict a label with high confidence.

- **Limitation:** Not applied in Real Time.

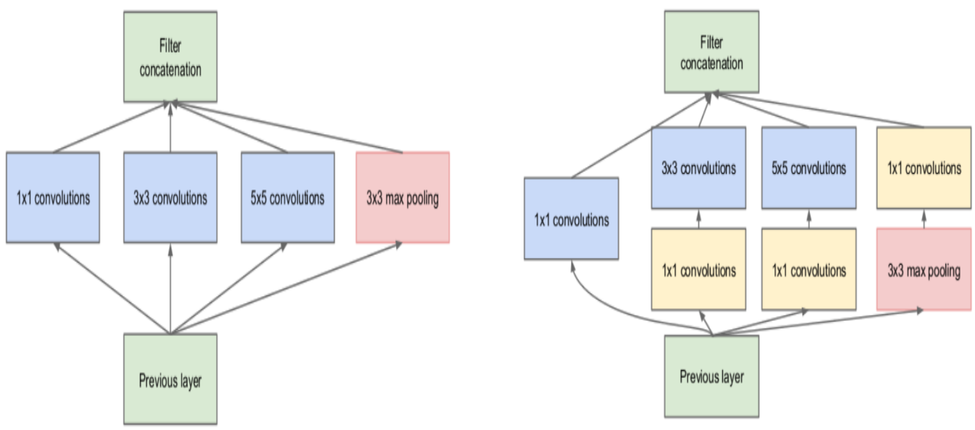

Figure 3: Naive Inception Module & Inception Module with Dimensionality Reduction

2.2.1.3 YOLO[10]:

- YOLO divides up the image into a grid of 13*13 cells. Each of the cell is responsible for predicting 5 bounding boxes.
- Each box prediction has 5 components (x,y,h,w,confidence). Confidence determines if the shape of the box is any good. For each bounding box, the cell also predicts a class.
- The confidence score and class prediction is combined to output one final score, if the box contains any object or not. The threshold is 30
- 3*13*5=845 boxes, all predicted at the same time, runs through the network only once. This ensures huge speed boost up.

- **Limitation:** To trade for speed, some amount of accuracy is sacrificed.

Figure 4: YOLO Methodology

2.2.1.4 R-CNN[11]:

- Generate a set of proposals for bounding boxes.
- Run the images in the bounding boxes through a pre-trained AlexNet and finally an SVM to see what object the image in the box is.
- Run the box through a linear regression model to output tighter coordinates for the box once the object has been classified.

- **Limitation:** R-CNN works well, but is quite slow for a few simple reasons:
  - It requires a forward pass of the CNN (AlexNet) for every single region proposal for every single image (that's around 2000 forward passes per image!).
  - It has to train three different models separately - the CNN to generate image features, the classifier that predicts the class, and the regression model to tighten the bounding boxes. This makes the pipeline extremely hard to train.

-

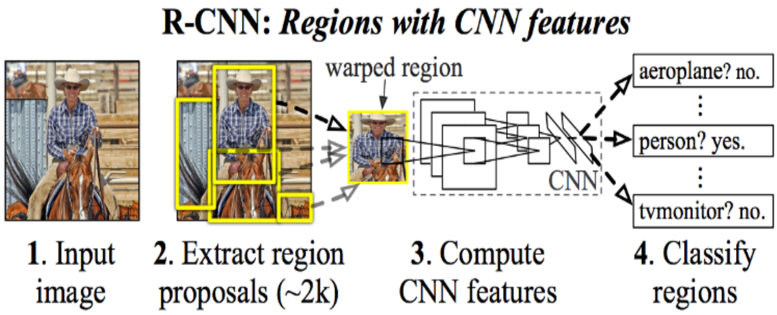

Figure 5: R-CNN Methodology

2.2.1.5 Fast R-CNN[12]:

- Fast R-CNN Insight 1: RoI (Region of Interest) Pooling For the forward pass of the CNN, Girshick realized that for each image, a lot of proposed regions for the image invariably overlapped causing us to run the same CNN computation again and again (2000 times!). His insight was simple Why not run the CNN just once per image and then find a way to share that computation across the 2000 proposals?

- Fast R-CNN Insight 2: Combine All Models into One Network: The second insight of Fast R-CNN is to jointly train the CNN, classifier, and bounding box regressor in a single model. Where earlier we had different models to extract image features (CNN), classify (SVM), and tighten bounding boxes (regressor), Fast R-CNN instead used a single network to compute all three.

- **Limitation:** Even with all these advancements, there was still one remaining bottleneck in the Fast R-CNN process, the region proposer.

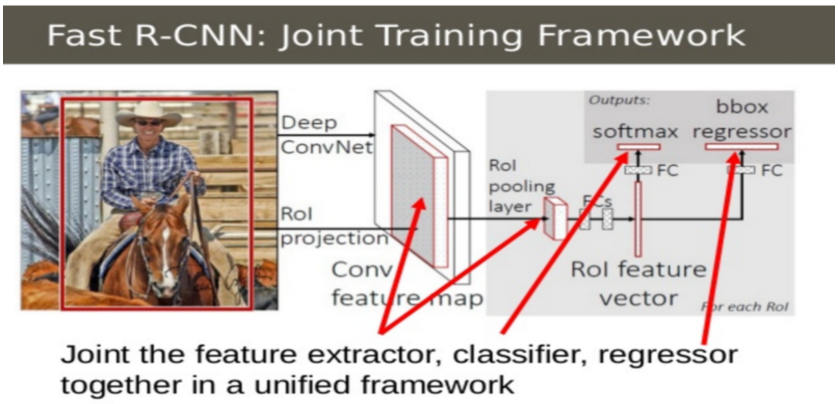

Figure 6: Fast R-CNN Methodology

2.2.1.6 Faster R-CNN [13]:

- The Region Proposal Network works by passing a sliding window over the CNN feature map and at each window, outputting k potential bounding boxes and scores for how good each of those boxes is expected to be. What do these k boxes represent?

- Intuitively, we know that objects in an image should fit certain common aspect ratios and sizes. For instance, we know that we want some rectangular boxes that resemble the shapes of humans. Likewise, we know we won't see many boxes that are very thin. In such a way, we create k such common aspect ratios we call anchor boxes. For each such anchor box, we output one bounding box and score per position in the image.

- With these anchor boxes in mind, let's look at the inputs and outputs to this Region Proposal Network: Inputs: CNN Feature Map. Outputs: A bounding box per anchor. A score representing how likely the image in that bounding box will be an object.

- **Limitation:** Accuracy is still a problem.

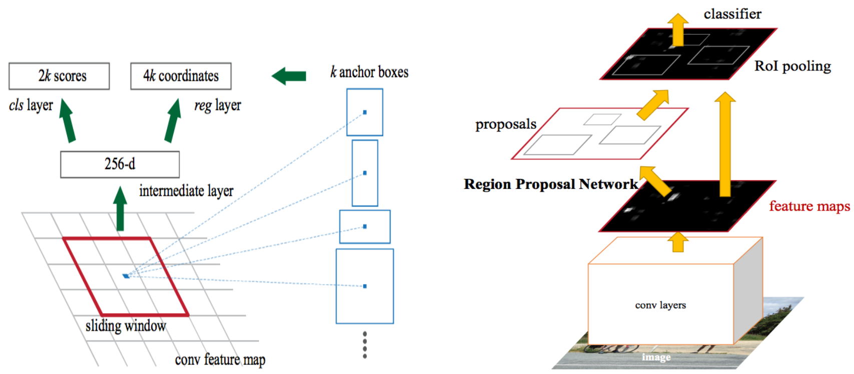

Figure 7: Faster R-CNN Methodology

2.2.2 Navigation Systems for the BVI

2.2.2.1 Without using Computer Vision:

In recent years, there has been quite some work on Navigation Assistance for Vsually Impaired (NAVI). In general, most of them do not require visual information for successful navigation rather a complex setup of hardware system. One such notable approach by MHA Wahab et al. [14] required a combination of multiple sensors including ultrasonic and water sensors, along with a micro controller. Another system designed by AS Al-Fahoum et al. [15] included multiple infrared sensors along with a PIC micro controller attached with a head hat and a minin hand stick. Approaches like these has been proved to be helpful for navigation while avoiding the obstacles on the way. As for the sensor based systems, the detection of obstacles might be accurate and efficient, but it is burdensome to classify them. Only a few categories at best can be classified based on the data given by the sensors.

2.2.2.2 Using Computer Vision:

More robust approaches in recent years are based on computer vision. As these systems are based on visual data, it is possible to provide more information to the subject than the approaches based primarily on sensor's data. Obstacle detection for the BVI people using computer vision is quite common in recent years. While the sensor based systems fails to classify obstacles of different categories, computer vision based systems excels in this criterion. Several approaches have been made to detect one or multiple specific type of objects from the visual data, to aid while navigation. One of such primary approaches was by X. Chen and A.L. Yuille [16], where they performed text detection from city scenes and read alound to the blind subjects based on a time efficient cascade system. Another such approach by Yingli Tian et al. [17], implemented the detection of doors from a known or unknown environments based on computer vision. They used cameras attached to a sunglass and a cap, they form a geometric door model, then detect certain features of door and decide it's position. On the other hand, Tess Winlock et al. [18] designed an advanced system that enables the BVI people to complete their shopping in groceries without any human assistance. They used their own detection algorithm ShelfScanner, which performs effectively given that there is sufficient training data provided.

2.2.2.3 Obstacle Detection Based:

There has also been notable works on computer vision based navigation and obstacle detection in both indoors and outdoors. Luis A. Guerrero et al. [19] designed a complete navigation system for the BVI people that works indoors. While Alberto Rodriguez et al. [20] designed an obstacle detection system that works in both indoors and outdoors. A stereo camera is carried by the user that is used calculate a dense disparity map based on the captured images. The map is then analyzed to detect potential obstacles and feedback is given to the subject via 'beep' sounds. A different approach was made by Ruxandra Tapu et al. [21], where they implemented obstacle detection via smart-phone devices. The region of interest is extracted from the captured images based on Lukas-Kanade [22] algorithm and perform further refinements. They categorized obstacles as urgent and normal based on their distances and feedbacks the user accordingly. There is comparatively far less hardware requirements in this approach. Another such system with minimum hardware requirements was designed by A. Aladren et al. [23] that requires only one RGB-D sensor with range expansion. Both of these systems are comparatively less complex in terms of hardware requirements than others.

2.2.2.4 Specific Object Detection Based:

Few approaches in recent years have been made to detect objects instead of obstacles for the BVI people. As the primary goal is to navigate safely, which can be done with obstacle detection alone. But in case of object detection systems, although they are troublesome to implement for several reasons, they can provide more information to the user than any other systems based on obstacle detection. The users are able to grasp an idea of the environment around them by knowing exactly what objects are around them, rather than simply classifying all of them as obstacles. Some obstacle detection can classify obstacles within a minimum number of classes [17] [18], but that is not quite enough to realize the total environment.

2.2.2.5 Object Detection Based:

Object Detection approaches to aid the BVI people are not quite much compared to obstacle detection approaches. One of the notable work was done by Hanen Jabnoun et al. [24], they implemented SIFT object detection algorithm to extract key points of the objects from video and finally detect the object. Another work by Hsueh-Cheng Wang [25], which primarily is an

NAVI system, performs object detection for certain objects like an empty chair etc. There are several reasons for not implementing object detection algorithms in the systems designed for the BVI people, we describe in detail in section 1.3 .

2.3 YOLO[10] for the NAVI system:

For our proposed system, we use YOLO [10] to perform the object detection task for several reasons -

1. YOLO object detection algorithm is comparatively faster than any object detection algorithm of recent years. As, we have to perform detection in real time, speed is a must, where YOLO is a reasonable preference.

2. As the system is actually a NAVI system, so portability has to be ensured. And for high configuration devices, it is difficult to ensure portability. Again, most of the object detection algorithms need high end setup to perform properly. So, as we trade off high end devices for portability, our speed of detection decreases significantly on mid end devices. So, we need the fastest detection algorithm to minimize the fact. So, YOLO remains a reasonable option.

3. Although, YOLO is not the best detection algorithm in terms of accuracy. But as we combine obstacle detection with object detection, this lacking is mostly covered with convenience.

3 Proposed Method:

3.1 Skeleton of the Proposed Method:

In this work, we propose a complete system framework combining both object detection and obstacle detection that ensures both effectiveness and efficiency in real time. A depth sense camera [1] and a portable computing device [2] is used to complete the hardware requirements. As the hardware components initialized our system starts. Now, as our system framework combine two types of detection, we the steps of object detection first, then the steps of obstacle detection. Finally, we discuss how the two are combined.

The major steps goes as follows:

1. **Input:** The D435 camera captures both RGB images and depth matrixes in real time with approximately 10-12 frames per second.
2. **Preprocessing:** The noises from the RGB images are reduced and unnecessary data from the depth matrix are excluded.
    (a) **Obstacle Detection:** If the minimum distance count is within urgent distance, we perform obstacle detection right away.

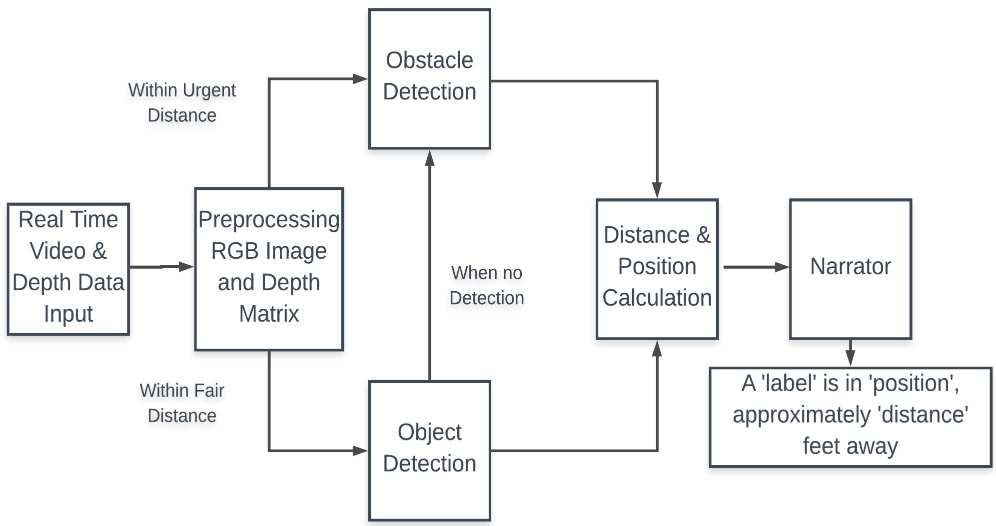

Figure 8: Skeleton of the Overall Methodology

(b) **Object Detection:** If the minimum distance count is not within urgent distance, we first perform object detection, if any object is detected then we forward to next step, else we switch to obstacle detection.

3. **Distance and Position Calculation:** We calculate the distance and position of the detected object/obstacle relative to the user following our own algorithm.

4. **Narrator:** Narrator combines the class label following the object detection, if obstacle detection is performed instead, this label simply assigned as 'obstacle'. Distance and Position comes from previous steps.

5. **Output:** Output is a string narrated to the subject via any hearing device. Sample output: ' A "person" is in "mid-right" , approximately "six" feet away'.

3.2 Object Detection:

The steps of object detection are as follows:

1. **Determine Indoor or Outdoor:** As we design our navigation system for both indoors and outdoors, there are certain differences in both environments that has to be taken into account. For indoors, the objects tend to be closer than outdoors. Again, moving objects are also less expected indoors compared to outdoors. Based on this, we fix three distances prior to our system start initialization: Min, Urgent and Max. They stay unchanged throughout the system performs. We define them as-

   - **Min-Distance**: Any object or obstacle within this distance is ignored. As the user's body parts, clothing are likely to be in front of the camera lens within this much close distance and these are not needed to be detected. For both indoors and outdoors, we fix this distance to be one feet.
   - **Urgent-Distance**: Any object or obstacle within this distance means it is very close to the user and it has to be detected right away. As object detection time is comparatively longer than obstacle detection time, when anything gets detected within this distance, it will be detected as an obstacle. Which means, no object detection within Urgent distance.

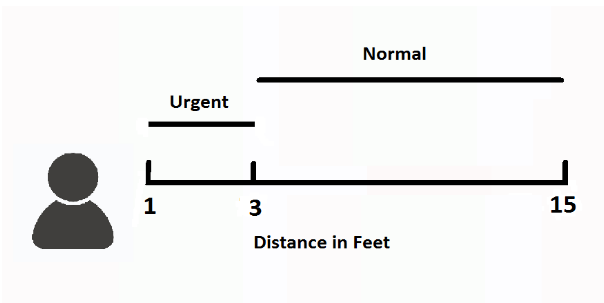

Figure 9: Indoor Distance Metrics

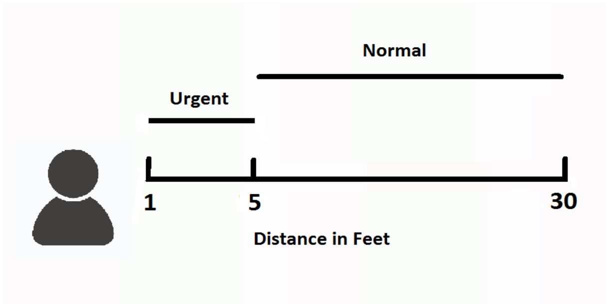

Figure 10: Outdoor Distance Metrics

For indoors, as objects are likely to appear close to the user, we set this distance to be three feet. For outdoors, where objects are tend to be more distant, we set this distance to be five feet.

- **Max-Distance**: Any object or obstacle that is quite far away from the BVI people is normally not their concern. Rather trading off detecting distant objects with detecting objects within concerned range is the prior choice. For indoors, this distance is fixed to 15 feet and for outdoors it is fixed to 30 feet.

From Urgent to Max Distance, we perform both detection. Firstly object detection, if the detection algorithm detects no object, then perform obstacle detection.

So, to summarize -

Table I
Detection with Distance

Distance	Detection
0 - Min	Ignore
Min - Urgent	Obstacle
Urgent - Max	Obstacle and Object
Max - Inf	Ignore

2. **Initialize Camera:** After determining indoor or outdoor environments and the distance metrics accordingly, we initialize the depth camera along with the computing device. It can be either an up board that ensures maximum portability or any laptop computing device which will also ensure sufficient portability. The camera is connected to the computing device via USB-C port and starts taking input.

3. **Capture Depth Image and RGB Image:**

   - **Capture RGB Image:** RGB images are captured with 6 FPS and 1280*720 size. Higher rate reduces the detector performance.
   - **Capture Depth Image:** Depth images are captured in parallel with RGB images. Every RGB image is associated with one unique depth image.

4. **Optimize Depth Matrix and Reduce Noise of RGB Image:**

   - **Optimize Depth Matrix:** Depth Matrix contain the depth data of it's associated RGB image. Both Depth Matrix and the RGB image are of same size and for each pixel of the RGB image, the depth matrix holds its corresponding distance. While optimizing, we exclude all distance values less than Min distance and greater than Max distance from the depth matrix. We assign a value greater than Max to all these, now Zero to Min and Max to Infinity values are ignored. The Depth Matrix Optimization (DMO) is performed as algorithm **??**
   - **Reduce Noise of RGB Image:** Linear smoothing filters are used to reduce noise from the RGB imsge.

5. **Object Detection:** The processed RGB image is given as input to the YOLO [10] detection system. YOLO performs object detection on the whole image and if any object is detected, it generates a bounding box and returns (x,y,w,h,label) for that object. (x,y) is the lower left corner co-ordinate of the bounding box and (w,h) is the width and height of the box, while label is the class name of the detected object. After the completion of object detection, (x,y,w,h) are passed to distance point calculation and label is passed to the narrator.

6. **Calculate Distance** At this stage, we have the optimized depth matrix and (x,y,w,h) values from YOLO. We get the bounding box and

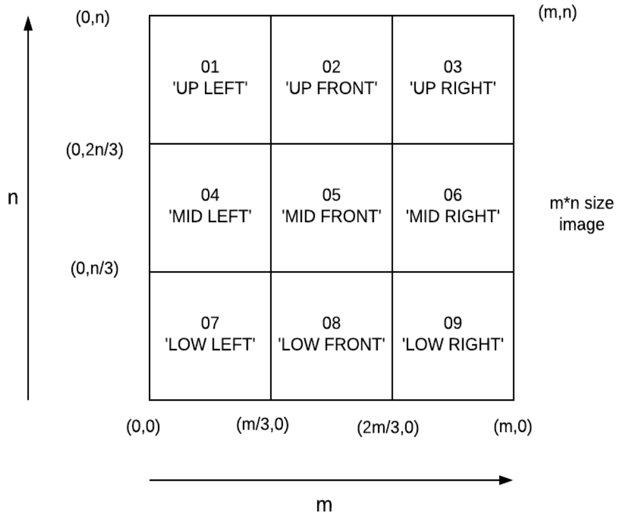

Figure 11: Region Mapping

place it on the depth matric. We take the mid point of the bounding box (x+w/2 , y+h/2). The corresponding distance value of this mid point is taken as the distance of the detected object from the user. Now, this calculated object distance is passed to the narrator and the calculated object midpoint is passed to calculate the position of the object relative to the whole image.

7. **Calculate Position** We divide the whole image into nine regions as given in figure 11. Then calculate on which region the point fells. That region name is taken as the position of the object. Algorithm 1 describes how it is calculated. The Position Calculation (PC) algorithm we propose to detect one of the nine regions given in figure 11, takes the midpoint of the object(x,y) and the RGB image as input. It returns one integer value from one to nine. The region name associated with it is the position of the object. The position is then forwarded to the narrator.

8. **Narrator** We keep a predefined string to give as audio output to the user with three variable parts that is replaced with data from the previous steps. The predefined string is: ' A "label" is in "position", approximately "Object Distance" feet away' . The portions bounded with double quotation marks are replaced according to the calculated

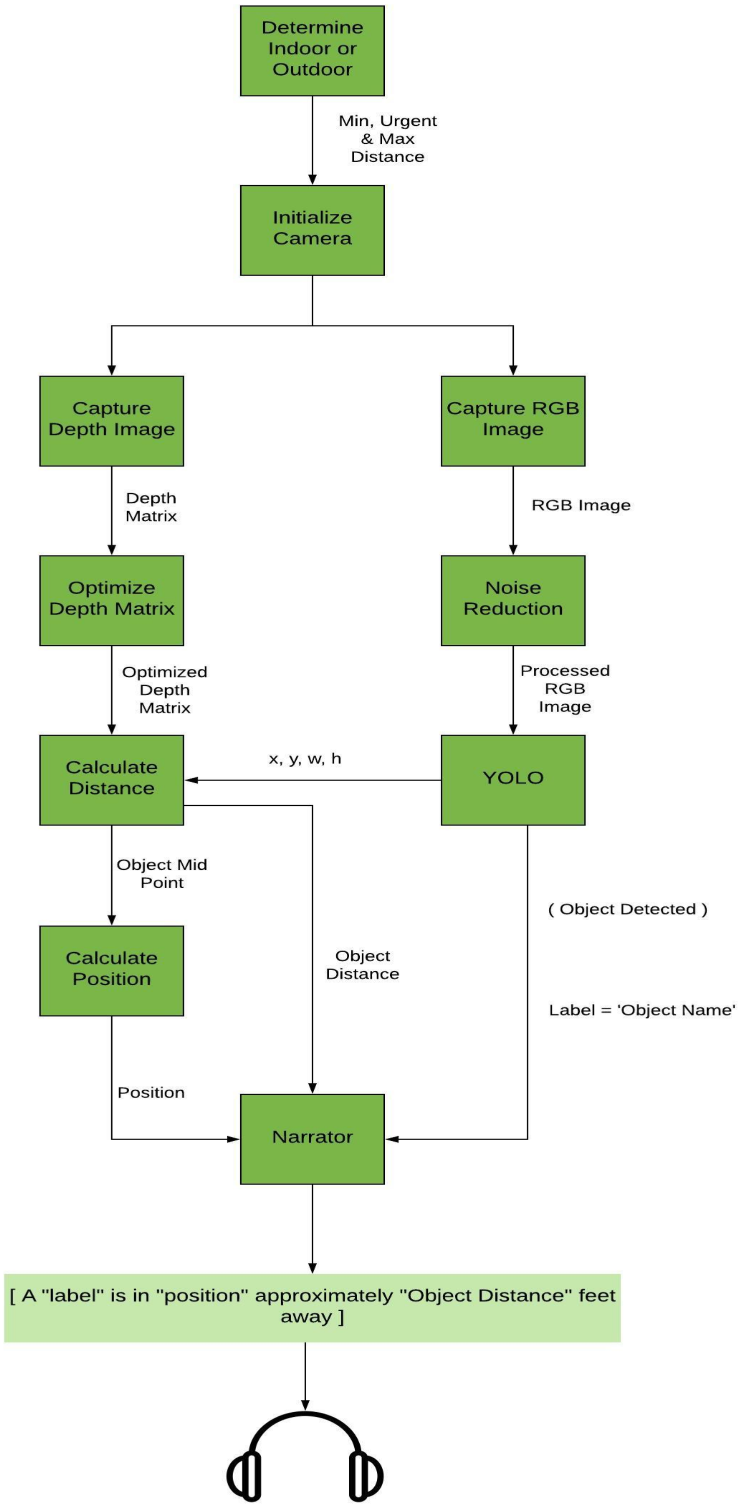

Figure 12: Object Detection Framework

Algorithm 1 Position_Calculation

```
1: procedure PC(x,y,RGB)
2: (m, n) ← size(RGB)
3: xx ← ceiling((x ∗ 3)/m)
4: yy ← ceiling((y ∗ 3)/n)
5: position ← xx + 3 ∗ (yy − 1)
```

data. From YOLO, "label" or class name is provided. From distance calculation, "Object Distance" is provided. And from position calculation, the "position" or one of the ninth region names of figure 11 is provided. As for an example - if YOLO detects a person in the processed RGB image. the distance calculation gives the distance of the mid point as five feet and the mid point falls in region '04', the output string will be: ' A *person* is in *mid-left*, approximately *five-feet* away '. We perform the text to speech operation using pyttsx3 [26], a Python, Open Source, text-to-speech library, for cross platforms.

9. **Headphone/Earphone:** Any normal earphone or headphone will work just fine. The narrator using pyttsx will convert the instruction into audio. With the assistance of earphone or headphone it will be delivered to the user.

3.3 Obstacle Detection

Although there has been a revolution in the field of object detection regarding higher and higher accuracy with time, but for obvious reasons no object detection ensures 100% accuracy. We present a chart comparing the accuracies of the recent detection algorithms in figure 13. This means, not every time, from every frame extracted from any video stream, there will be detected objects. The detection algorithm might not detect any objects from the input image due to lack of enough data about the object or simply there might not be any object to detect in the image. But there still might be some obstacles in the pathway, again any undetected object remains an obstacle. This scenario poses a huge challenge in implementing object detection algorithms for NAVI systems. We back this up by simply combining obstacle detection with generic object detection. We perform obstacle detection in two scenarios:

In the system framework, the steps of obstacle detection are identical to the steps of object detection for the first scenario mentioned above. Only there are some slight modifications, as shown in figure 14. The modifications are as follows:

1. Steps one to four remains exactly the same.
2. In step five, YOLO returns NULL. So instead of the class name, the label is set to 'obstacle' and passed to the Narrator. And no data is sent from YOLO for distance calculation like before, as in figure 12.

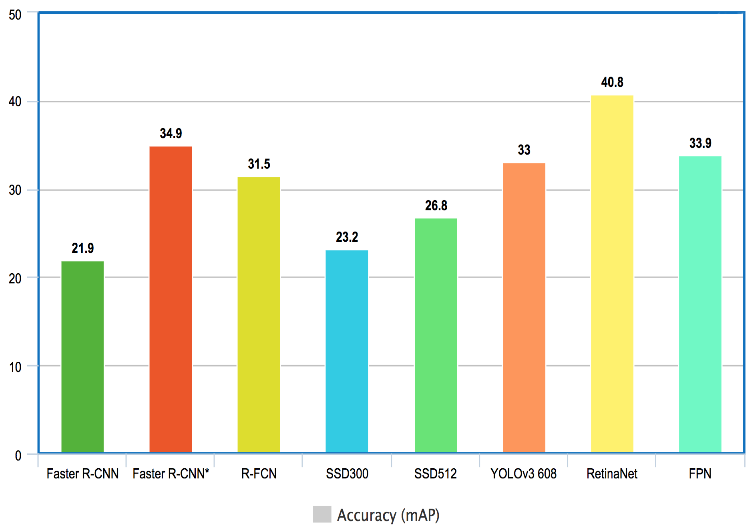

Figure 13: Accuracy Comparison of Object Detection Algorithms

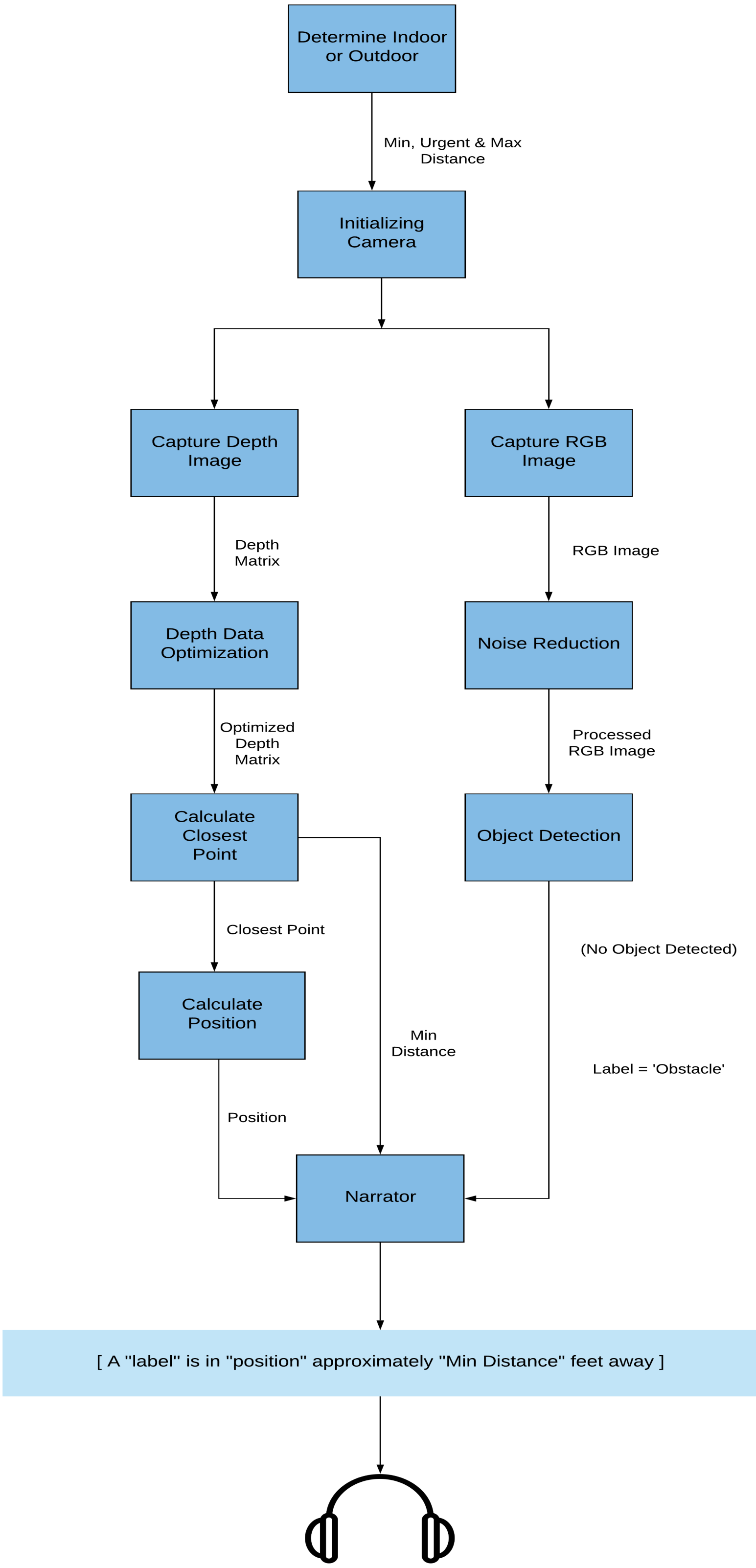

Figure 14: Obstacle Detection Framework

3. In step six, no mid point is calculated as there is no object and it's bounding box. Here we have only the optimized depth matrix. To detect obstacle, we just calculate the minimum distance from the matrix. The reason behind choosing the minimum distance point is ensuring collision avoidance. The co-ordinate of the minimum distance point is then passed to the narrator and also to calculate position.

4. Step seven requires some modification. As, in case of object detection, we calculated the position of the detected object's mid point, but in case of obstacle detection, we just calculate the closest point.

5. Step eight is almost same, only in the place of "label", there will always be "obstacle".

6. Step nine remains unchanged also.

3.4 Combining Object Detection with Obstacle Detection

In this section, we describe how we combine both detection systems simultaneously in real time. We explain the total framework design hierarchically step by step below:

1. Step one to four, which is from determining indoor or outdoor to generating optimized depth matrix and processed RGB image remains same.

2. As soon as we get the optimized depth matrix, we calculate the closest point from the distance values.

3. If the closest point is within urgent distance, we perform obstacle detection right away.

   (a) The closest point distance is sent to the narrator and also to calculate the position.

   (b) Upon completion of position calculation, the position is also sent to the narrator.

   (c) Narrator now have the closest distance point and it's position. As, no object detection is involved, narrator will define it as an obstacle and give the audio output as we described in section **??**.

   (d) After the narrator being done, the system again will again perform step one to four.

   (e) If the closest point this time is not within urgent distance, we stall the system for five seconds.

   (f) If the closest point this time is within urgent distance, we perform obstacle detection immediately based on the newly calculated closest point.

   Here, we normally wait for five seconds after any kind of detection, but if any object is found within urgent range we perform obstacle detection right away. But if not urgent, continuous audio instruction can be a bother to the BVI subject. Again, after receiving one piece of information, they should be given some time to process it. Hence the wait of five seconds, unless the scenario of urgent.

4. If the closest point is not within urgent distance, we pass the processed RGB image to YOLO algorithm.

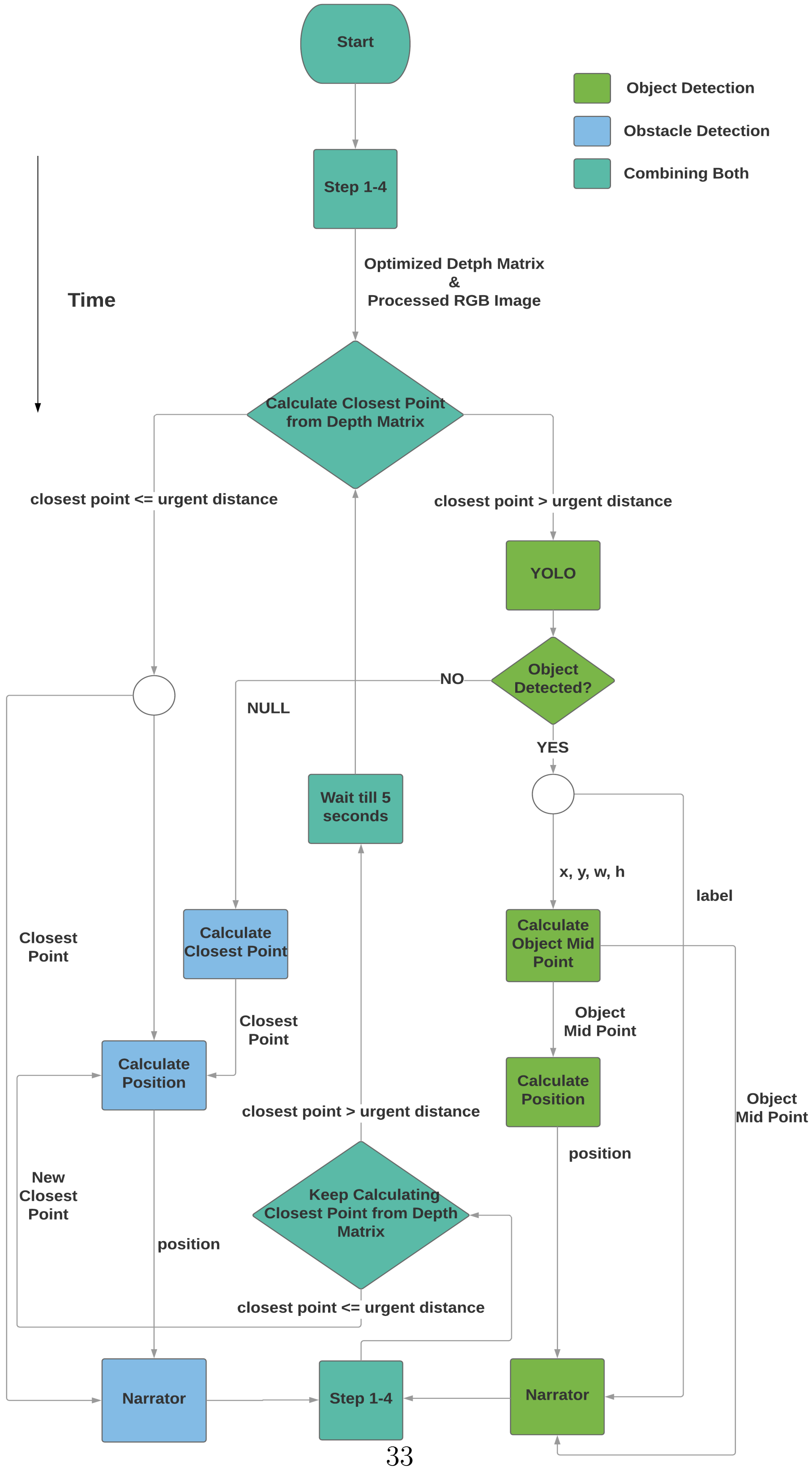

Figure 15: Object Detection Framework

5. Now, if any object is detected,

    (a) The values regarding the bounding box (x,y,w,h) are passed to calculate the object mid point and label of the classified object is passed to the narrator.
    (b) After the calculation of mid point it is passed to calculate the position and also to the narrator.
    (c) The calculated position is also passed to the narrator.
    (d) After the narrator is done giving the audio output to the subject, steps one to four will be performed again.
    (e) Now, just like obstacle detection, we wait for five seconds, if the newly calculated closest point is not within urgent range.
    (f) Otherwise, we pass the newly calculated closest point to perform obstacle detection like before.

6. If no object is detected,

    (a) So, YOLO returns NULL here. We perform the calculation of closest point based on the depth matrix only. We pass this point to obstacle detection.
    (b) We perform obstacle detection like before, by calculating position and passing it to the narrator.
    (c) Again, we wait for five seconds, if the new closest point is outside of the urgent range.
    (d) Otherwise, we switch to obstacle detection immediately.

4 Performance Evaluation

4.1 Dataset and Experimental Setup:

We performed our experiment on a Laptop Computer, with 8 GB of RAM, NVIDIA 940 MX graphics card, Core i5 2.50 GHz Processor and CUDA environment setup.

YOLO is trained on COCO dataset [27], that has 80 different class labels. So, upto 80 different classes can be detected by our system and other objects or undetected objects will be taken as obstacles. We perform our combined object and obstacle detection in both indoors and outdoors in real time and finally gather the data to calculate performance metrics. We also perform our experiment in different times of the day, also in different lighting conditions.

Table 1: Sample Frames from Video Inputs

We have taken multiple real time video inputs with both potential objects and obstacles in it. We determine the number of objects present in the video within urgent to max distance manually and compare with the number of objects the system could actually detect.

4.2 Performance Measurement:

The system is tested in multiple different environments for short period of real time video input data in both indoor and outdoor with specific number of objects. The res

Table II
Indoor Object Detection Results

Video	Total Number of Frames	Total Number of Objects	Detected Number of Objects	%
1	43	5	4	80 %
2	74	2	2	100 %
3	255	9	6	66.66 %

Table III
Outdoor Object Detection Results

Video	Total Number of Frames	Total Number of Objects	Detected Number of Objects	%
4	27	15	12	80 %
5	198	21	17	81.3 %
6	121	35	23	65.7 %
7	249	41	34	82.9 %

We have measured the performance by calculating Precision and Recall. Precision (also called Positive Predictive Value) is the fraction of retrieved instances that are relevant; While Recall (also called sensitivity) is the fraction of relevant instances that are retrieved. High recall means that an algorithm returned most of the relevant results, while high precision means that an algorithm returned substantially more relevant results than irrelevant. For classification tasks, the terms true positives, true negatives, false positives, and false negatives compare the results of the classifier under test with trusted external judgments.

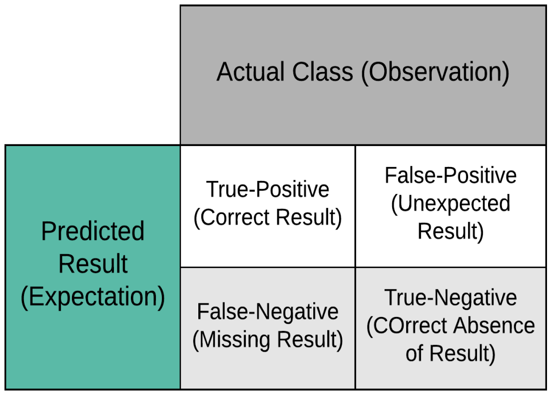

Figure 16: Object Detection Framework

The terms positive and negative refer to the classifier's prediction (sometimes known as the expectation), and the terms true and false refer to whether that prediction corresponds to the external judgment (sometimes known as the observation). This is illustrated by figure 16.

For object detection and obstacle detection our result sums up to -

Table IV
Confusion Matrix for Object Detection

	True	False
Positive	71.2%	5.36%
Negative	18.2%	5.3%

Table V
Confusion Matrix for Obstacle Detection

	True	False
Positive	91.1%	4.8%
Negative	3.9%	0.2%

4.3 Comparative Analysis

Dakopoulos et al. [28] conducted a survey over Wearable Obstacle Avoidance Electronic Travel Aids for Blind. They compared all notable NAVI tools based on some major structural and functional features. Every feature had separate weight values based on their importance and necessity. Users tested the systems and gave scores out of 10 for every features and then the average feature was multiplied by their specific weight value. Finally, all the result values are summed up to produce a score within 50.

The structural and functional features along with their respective weights are as follows -

Table VI
Structural and Functional Features of a NAVI System[28]

#	Feature	Weight
F1	Realtime	9.3
F2	Wearable	8.6
F3	Portable	5.7
F4	Reliable	7.1
F5	Low-Cost	5.0
F6	Friendly	4.3
F7	Functionalities	2.7
F8	Simple	2.9
F9	Robust	2.1
F10	Wireless	1.4
F11	Performance	10
F12	Originality	1.4
F13	Availability	5.0
F14	Future	6.4

We conducted our experiment on four subjects and gathered their scores for each features. Multiply the scores with weights and sum them up. If we do not have enough information for a feature of a system, we do not assign any score. Note that in this evaluation we provide for availability of the device, and its wireless feature the value 10 or no value for computational reasons.

$$S = \sum_{i=1}^{N} \frac{w_i x_i}{N} + 2$$

where i refers to a specific feature, N is the total number of features for each system, and b is bias (for now, b = 2).

S is the maturity score for every system.

We conducted our experiment on four individuals and let them evaluate the scores for the specific and functional features, through which we finally calculated S_i for the system. Gathering all four, we took the average maturity score and compare it with all others.

Table VII
Subject Feedback

Subject #	Maturity Score S_i
01	45.47
02	47.82
03	47.50
04	49.01

Finally, Maturity Score for our system, $S_{avg} = 47.45$

Now we compare this score with the all existing Commercial , notable NAVI systems with the same evaluation method:

Table VIII
Maturity Score Comparison with Commercial Systems

#	System Name	S
A	Echolation [29]	35.2
B	Navbelt [30]	32.1
C	vOICe [31]	37.7
D	University of Stuttgart[32]	36.3
E	FIU [33]	40.1
F	Virtual Acoustic Space [34]	40.2
G	NAVI [35]	41.5
H	University of Guelph [36]	40.9
I	GuideCane [37]	35.0
J	ENVS [38]	44.2
K	CyARM [39]	36.6
L	Tactile Handle [40]	36.0
M	TVS [42]	44.0
N	**EPFL** [43]	**46.1**
O	Tyflos [44]	43.2
P	FIU cv project [45]	29.3
Q	UCSC [46]	32.2
R	**Proposed System**	**47.45**

The maturity ranking Table VIII gives us a big picture for all the reviewed NAVIs; a measure of the system's progress/maturity. The ones with higher scores show better progress and/or more features. The systems that got lower scores are not of less technological or usage value, but they are still in the early stage of their progress and they have not reached their maximum of their performance. Finally, our proposed system showed better overall performances than all other existing systems.

4.4 Output and Result Images

Object detection results in different environments along with their depth matrix are as follows:

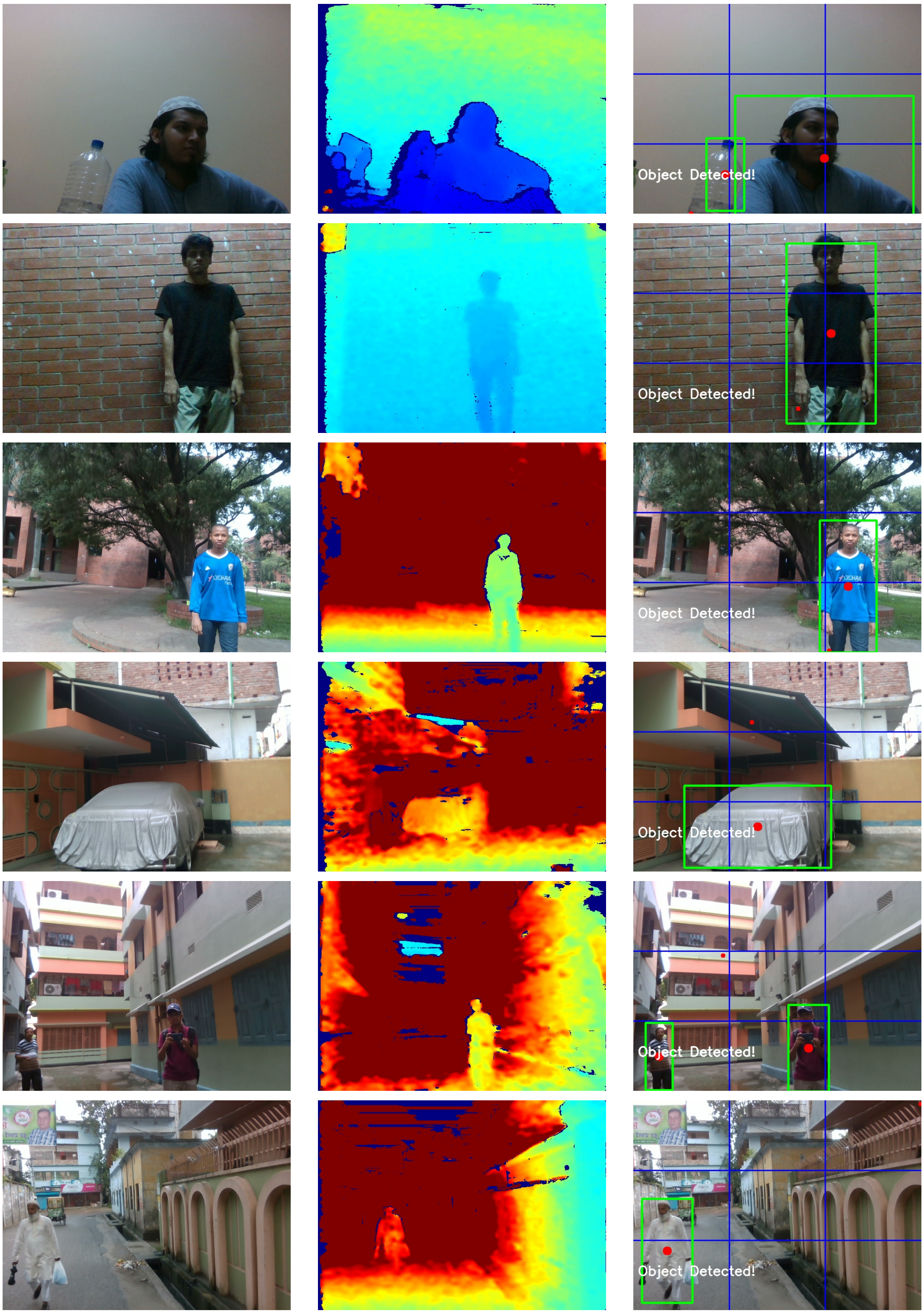

Figure 17: (From Left to Right) Original RGB Image, Depth Image and Detected Objects with Bounding Boxes

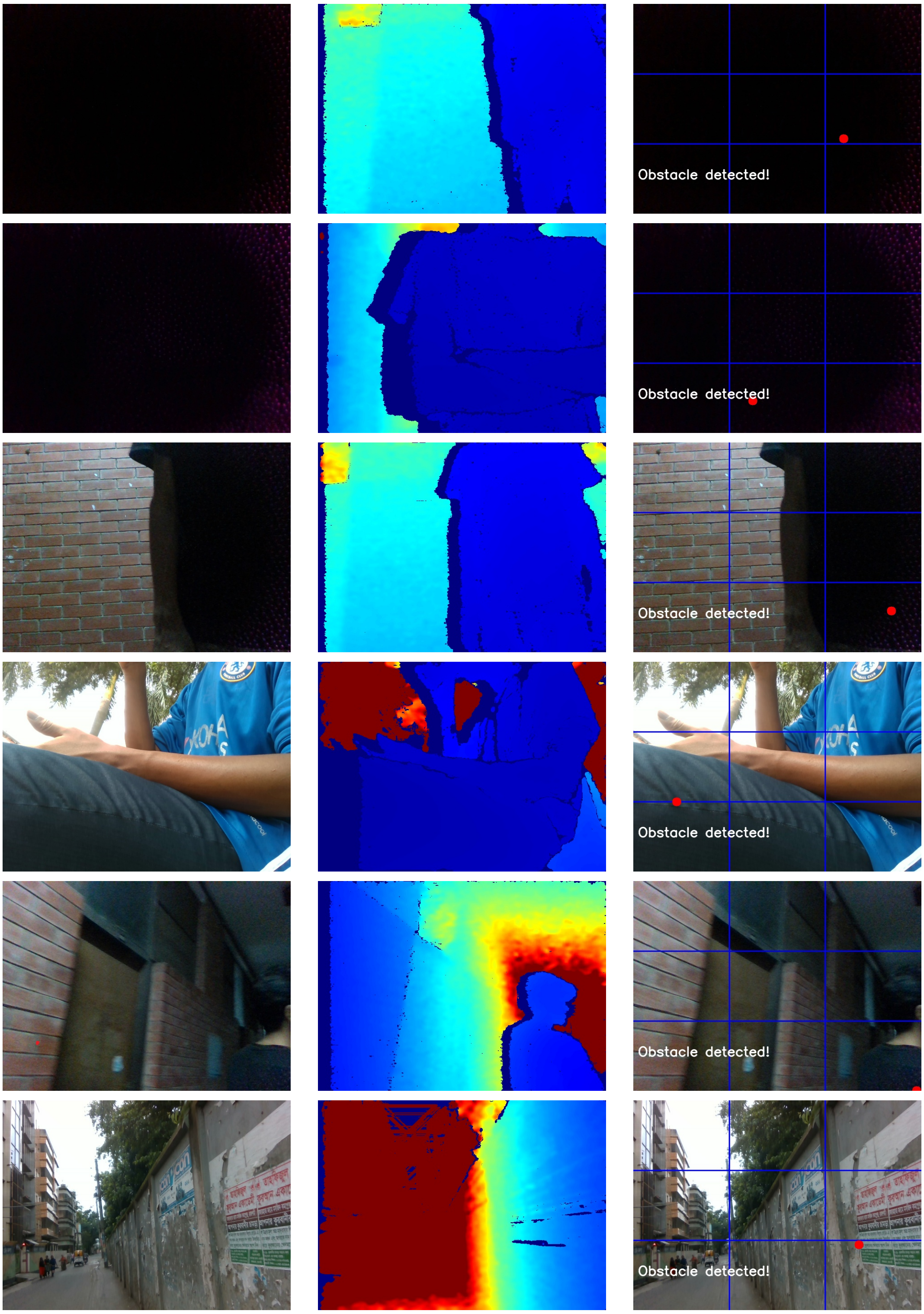

Figure 18: (From Left to Right) Original RGB Image, Depth Image and Detected Obstacle with Closest Point. First Two Row Demonstrates Obstacle Detection in the Dark.

5 Conclusion and Future Works

The system we proposed performs effectively in different environments both in indoor and outdoor. As we implemented our work via a Laptop device, portability is ensured, but with the use of lighter computing devices as we theoretically proposed with Up Board [2] will ensure far more portability and robustness. Again, another issue remains with the system being a little costly. Considering all the factors and performances, the cost might be reduced in future. We also intend to work further on simplicity and convenience of the system.

6 Appendix

6.1 Subject Information

1. Name: Md. Abdul Bari
   Age : 20
   Visual Impairment: Right Eye: -2.75 Left Eye: -3.15

2. Name: Alsaad Ahmed
   Age : 23
   Visual Impairment: Right Eye: -3.75 Left Eye: -3.50

3. Name: Muhtasim Jawad Nafi
   Age : 21
   Visual Impairment: Right Eye: -3.00 Left Eye: -2.75

4. Name: Ishtiaque Ahmed Sonnet
   Age : 19
   Visual Impairment: Right Eye: -4.75 Left Eye: -4.50

References

[1] Intel RealSense Camera D435, Intel. https://click.intel.com/intelr-realsensetm-depth-camera-d435.html

[2] Intel Up Board, Intel. https://www.up-board.org/

[3] Pradeep, V.; Medioni, G.; Weiland, J. Robot Vision for the Visually Impaired. In Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 15–22.

[4] Alcantarilla, P.F.; Yebes, J.J.; Almazan, J.; Bergasa, L.M. On Combining Visual SLAM and Dense Scene Flow to Increase the Robustness of Localization and Mapping in Dynamic Environments. In Proceedings of the IEEE International Conference on Robotics and Automation, Saint Paul, MN, USA, 14–18 May 2012; pp. 1290–1297.

[5] Dakopoulos, D.; Bourbakis, N.G. Wearable obstacle avoidance electronic travel aids for blind: A survey. IEEE Trans. Syst. Man Cybern. C 2010, 40, 25–35.

[6] Saez, J.M.; Escolano, F.; Penalver, A. First Steps towards Stereo-Based SLAM for the Visually Impaired. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Diego, CA, USA, 25 June 2005; p. 23.

[7] Saez, J.M.; Escolano, F. Stereo-Based Aerial Obstacle Detection for the Visually Impaired. In Proceedings of Workshop on Computer Vision Applications for the Visually Impaired, Marselle, France, 18 October 2008.

[8] Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[9] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition(pp. 1-9).

[10] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A., 2016. You only look once: Unified, real-time object detection. In Proceedings of the

IEEE conference on computer vision and pattern recognition (pp. 779-788).

[11] Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[12] Girshick, R., 2015. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[13] Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[14] Wahab, M.H.A., Talib, A.A., Kadir, H.A., Johari, A., Noraziah, A., Sidek, R.M. and Mutalib, A.A., 2011. Smart cane: Assistive cane for visually-impaired people. arXiv preprint arXiv:1110.5156.

[15] Al-Fahoum, A.S., Al-Hmoud, H.B. and Al-Fraihat, A.A., 2013. A smart infrared microcontroller-based blind guidance system. Active and Passive Electronic Components, 2013.

[16] Chen, X. and Yuille, A.L., 2005, June. A time-efficient cascade for real-time object detection: With applications for the visually impaired. In Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on (pp. 28-28). IEEE.

[17] Tian, Y., Yang, X. and Arditi, A., 2010, July. Computer vision-based door detection for accessibility of unfamiliar environments to blind persons. In International Conference on Computers for Handicapped Persons (pp. 263-270). Springer, Berlin, Heidelberg.

[18] Winlock, T., Christiansen, E. and Belongie, S., 2010, June. Toward real-time grocery detection for the visually impaired. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on (pp. 49-56). IEEE.

[19] Guerrero, L.A., Vasquez, F. and Ochoa, S.F., 2012. An indoor navigation system for the visually impaired. Sensors, 12(6), pp.8236-8258.

[20] Rodríguez, A., Yebes, J.J., Alcantarilla, P.F., Bergasa, L.M., Almazán, J. and Cela, A., 2012. Assisting the visually impaired: obstacle detection and warning system by acoustic feedback. Sensors, 12(12), pp.17476-17496.

[21] Tapu, R., Mocanu, B., Bursuc, A. and Zaharia, T., 2013. A smartphone-based obstacle detection and classification system for assisting visually impaired people. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 444-451).

[22] B. Lucas, T. Kanade, An iterative technique of image registration and its application to stereo, ICAI, 1981.

[23] Aladren, A., López-Nicolás, G., Puig, L. and Guerrero, J.J., 2016. Navigation assistance for the visually impaired using RGB-D sensor with range expansion. IEEE Systems Journal, 10(3), pp.922-932.

[24] Jabnoun, H., Benzarti, F. and Amiri, H., 2015, December. Object detection and identification for blind people in video scene. In Intelligent Systems Design and Applications (ISDA), 2015 15th International Conference on (pp. 363-367). IEEE.

[25] Wang, H.C., Katzschmann, R.K., Teng, S., Araki, B., Giarré, L. and Rus, D., 2017, May. Enabling independent navigation for visually impaired people through a wearable vision-based feedback system. In 2017 IEEE international conference on robotics and automation (ICRA) (pp. 6533-6540). IEEE.

[26] Python Software Foundation, Open Source, Text-to-Speech library for cross platforms, https://pypi.org/project/pyttsx3/

[27] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., 2014, September. Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

[28] Dakopoulos, D. and Bourbakis, N.G., 2010. Wearable obstacle avoidance electronic travel aids for blind: a survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(1), pp.25-35.

[29] T. Ifukube, T. Sasaki, and C. Peng, "A blind mobility aid modeled after echolocation of bats," IEEE Trans. Biomed. Eng., vol. 38, no. 5, pp. 461– 465, May 1991

[30] S. Shoval, J. Borenstein, and Y. Koren, "Mobile robot obstacle avoidance in a computerized travel aid for the blind," in Proc. 1994 IEEE Robot. Autom. Conf., San Diego, CA, May 8–13, pp. 2023–2029

[31] P. B. L. Meijer. (1992, Feb.). An experimental system for auditory image representations. IEEE Trans. Biomed. Eng. [Online]. 39(2), pp. 112–121. Available: http://www.seeingwithsound.com/

[32] A. Hub, J. Diepstraten, and T. Ertl, "Design and development of an indoor navigation and object identification system for the blind," in Proc. ACM SIGACCESS Accessibility Computing, Sep. 2003/Jan. 2004, no. 77–78, pp. 147–152

[33] D. Aguerrevere, M. Choudhury, and A. Barreto, "Portable 3D sound / sonar navigation system for blind individuals," presented at the 2nd LACCEI Int. Latin Amer. Caribbean Conf. Eng. Technol. Miami, FL, Jun. 2–4 2004.

[34] J. L. Gonzalez-Mora, A. Rodr ´ ´ıguez-Hernandez, L. F. Rodr ´ ´ıguez-Ramos, L. D´ıaz-Saco, and N. Sosa. (2009, May 8). Development of a new space perception system for blind people, based on the creation of a virtual acoustic space. Tech. Rep. [Online]. Available: http://www.iac.es/proyect/eavi

[35] G. Sainarayanan, R. Nagarajan, and S. Yaacob, "Fuzzy image processing scheme for autonomous navigation of human blind," Appl. Softw. Comput., vol. 7, no. 1, pp. 257–264, Jan. 2007

[36] R. Audette, J. Balthazaar, C. Dunk, and J. Zelek, "A stereo-vision system for the visually impaired," Sch. Eng., Univ. Guelph, Guelph, ON, Canada, Tech. Rep. 2000-41x-1, 2000

[37] I. Ulrich and J. Borenstein, "The guidecane – applying mobile robot technologies to assist the visually impaired people," IEEE Trans. Syst., Man Cybern., A: Syst. Hum., vol. 31, no. 2, pp. 131–136, Mar. 2001.

[38] S. Meers and K. Ward, "A substitute vision system for providing 3D perception and GPS navigation via electro-tactile stimulation," presented at the 1st Int. Conf. Sens. Technol., Palmerston North, New Zealand, Nov. 21–23, 2005

[39] K. Ito, M. Okamoto, J. Akita, T. Ono, I. Gyobu, T. Tagaki, T. Hoshi, and Y. Mishima, "CyARM: An alternative aid device for blind persons," in Proc. CHI05, Portland, OR, Apr. 2–7, 2005, pp. 1483–1488.

[40] M. Bouzit, A. Chaibi, K. J. De Laurentis, and C. Mavroidis, "Tactile feedback navigation handle for the visually impaired," presented at the

[41] 2004 ASME Int. Mech. Eng. Congr. RDD Expo., Anaheim, CA, Nov. 13–19

[42] C. Shah, M. Bouzit, M. Youssef, and L. Vasquez, "Evaluation of RU-netra – tactile feedback navigation system for the visually impaired," in Proc. Int. Workshop Virtual Rehabil., New York, 2006, pp. 71–77.

[43] L. A. Johnson and C. M. Higgins, "A navigation aid for the blind using tactile-visual sensory substitution," in Proc. 28th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., New York, 2006, pp. 6298–6292

[44] S. Cardin, D. Thalmann, and F. Vexo, "A wearable system for mobility improvement of visually impaired people," Vis. Comput., vol. 23, no. 2, pp. 109–118, Jan. 2007

[45] N. G. Bourbakis and D. Kavraki, "Intelligent assistants for handicapped people's independence: Case study," in Proc. 1996 IEEE Int. Joint Symp. Intell. Syst., Nov. 4–5, pp. 337–344.

[46] D. Dakopoulos, S. K. Boddhu, and N. Bourbakis, "A 2D vibration array as an assistive device for visually impaired," in Proc. 7th IEEE Int. Conf. Bioinf. Bioeng., Boston, MA, Oct. 14–17, 2007, vol. 1, pp. 930–937.

[47] D. Dakopoulos and N. Bourbakis, "Preserving visual information in low resolution images during navigation for visually impaired," presented at the 1st Int. Conf. PErvasive Technol. Related Assist. Environ., Athens, Greece, Jul. 15–19, 2008.

[48] M. Adjouadi, "A man-machine vision interface for sensing the environment," J. Rehabil. Res. Develop., vol. 29, no. 2, pp. 57–56, 1992.

[49] D. Yuan and R. Manduchi, "A tool for range sensing and environment discovery for the blind," in Proc. 2004 Conf. Comput. Vis. Pattern Recogn. Workshop, Washington, DC, Jun. 27–Jul. 02, vol. 3, pp. 39–39.