Guiding and Navigation for the Blind using Deep Convolutional Neural Network Based Predictive Object Tracking

Sai Nikhil Alisetti, Swarnalatha Purushotham, Lav Mahtani

Abstract: ***Indoor and outdoor Navigation is a tough task for a visually impaired person and they would most of the time require human assistance. Existing solutions to this problem are in the form of smart canes and wearable. Both of which use sensors like on - board proximity and obstacle detection, as well as a haptic or auditory feedback system to warn the user of stationary or incoming obstacles so that they do not collide with any of them as they move. This approach has many drawbacks as it is not yet a stand - alone reliable device for the user to trust when navigating, and when frequently triggered in crowded areas, the feedback system will confuse the user with too many requests resulting in loss of actual information.***

Our Goal here is to create a Personalized assistant to the user, which they can interact naturally with their voice to mimic the aid of an actual human assistance while they are on the move. It works by using its object detection module with a high reliability training accuracy to detect the boundaries of objects in motion per frame and once the bounding box crosses the threshold accuracy, recognize the object in the box and pass the information to the system core, where it verifies if the information needed to be passed onto the user or not, if yes it passes the converted speech information to the voice interaction model. The voice interaction model is consent-based, it would accept and respond to navigation queries from the user and will intelligently inform them about the obstacle which needs to be avoided. This ensures only the essential information in the form of voice requests is passed onto the user, which they can use to navigate and also interact with the assistant for more information.

Keywords: Vision Processing, Medical Aid, Voice Assistant, Real Time Object Detection, YOLO2000 Model

I. INTRODUCTION

General Navigation and Commuting is a rather tough ask fora person who has a partial or total loss of vision which cannot be medically corrected. They would almost always require human assistance for general everyday tasks which requires them to travel. There have been many solutions proposed to mitigate this problem using varied architectures, the stable and state of the art techniques include smart cane and wearable.

Both use a camera, a few sensors and on - board proximity and obstacle detection, as well as a haptic or auditory feedback system to warn the user of stationary or incoming obstacles so that they do not collide with any of them as they move. This approach has many drawbacks to itself as it is not yet a stand - alone reliable device for the user to trust when navigating, and when frequently triggered in crowded areas, the feedback system will confuse the user to the point where it defeats its purpose.

II. LITERATURE SURVEY

S. Ray et.al (2012) [1] in their paper "Detection, Recognition and Tracking of Moving Objects from Real-time Video via SP Theory of Intelligence and Species Inspired PSO" explore the confluence of object detection and natural language processing. In the experiment, the authors use natural language texts of multiple alignments for SP systems to extract necessary information from raw data line in form of old pattern and comprehend the knowledge base in test domain to derive and encode new information. This approach might be useful for optimized object detection. Anwar et.al (2017) [2] design a basic version of a smart stick for the visually impaired described in the paper "A smart stick for Assisting Blind People". Their approach is sensor based and uses buzzers yet it is serves as a good starting point for integrating multiple systems. They use Arduinos, ultrasonic sensors and heat sensors to guide blind people through hyperlocal environments. Their prototype is one step higher to the one introduced by J.Na which does not have any real time input feeds. We aim to have the author's solution as one part of our stick module which also, as previously mentioned, has NLP and object detection as its main features.

The insightful paper "Improving Video Activity Recognition using Object Recognition and Text Mining authored by Mooney et.al (2012) [3] do significant groundwork to show how object recognition in general is meshed with NLP. Using pre-processed visual words, a set of spatial temporal interest points are extracted from a video clip. To gather information on the correlation between activities and objects they mined the 2005 English Gigaword corpus. They use five distinct international sources of English newswire containing a total of 6 million documents or 15 GB of raw text. Using this corpus, they computed occurrence and co-occurrence counts for activities and objects.

Revised Manuscript Received on December 16, 2019.

***** Correspondence Author, [1] First Author

Sai Nikhil Alisetti[1], B.Tech, School of Computer Science and Engineering, Vellore Institute of Technology (VIT), Vellore, Tamil Nadu. Email: sainikhil.ds123@gmail.com

Swarnalatha Purushotham*, Associate Professor, School of Computer Science and Engineering, Vellore Institute of Technology (VIT), Vellore, Tamil Nadu.
Email: pswarnalatha@vit.ac.in

Lav Mahtani, B.Tech, School of Computer Science and Engineering, Vellore Institute of Technology (VIT), Vellore, Tamil Nadu.
Email: lav.mahtani@gmail.com

An occurrence of an activity Ai was defined as an occurrence (after stemming) of any one of the verbs in the verb cluster defining Ai. An occurrence of an object Oi was defined as any occurrence (after stemming) of the noun defining Oi or a "synonym." The set of synonyms was found by considering all nouns in the descriptive sentences in the training data and keeping those whose less similarity with the defining noun was greater than 0.5. This approach is domain-independent and hence can also be applied to the problem at hand.

Kocaleva et.al (2016) [4] through their paper "Pattern Recognition and Natural Language Processing: State of the Art" show how pattern recognition and natural language processing are interleaved with each other. The authors shed light on statistical NLP which bases NLP on a probabilistic definition so as to introduce estimation in the prediction of the interpretation of multi-words. This approach is useful to help us in estimating the accuracy of our predictions of what the objects around the blind person are. Yang et.al (2013) [5] use a similar approach for "multi-label visual recognition using NLP". They study robotics' dependencies on NLP. They in fact use two ways to show the importance of NLP. First, they use pre-defined databases that organize words according to semantics that are then used to predict activities. Second is to use correlation factors by building a knowledge base of daily activities and then exploiting correlations between these activities. The authors tested this approach on three real world tasks and achieved sufficient performance in all three of them.

Agarwal et al. [6] in their paper titled "Ultrasonic Stick for Blind" have worked on an economical solution, specifically an ultrasonic stick for the visually impaired. The device they developed consists of a user-friendly interface, with ultra-sonic sensors and point cameras. These Ultrasonic sensors have been placed in a way that they can scan three directions (at 180 degrees). The camera has been placed such that it can be used as an alternative tool in areas surrounded by low signal coverage. These are then paired with a microcontroller, a buzzer and a vibrating engine. When any obstacle is detected, the buzzer and vibration motor are activated. GPS system provides its current location information. The blind uses the SMS system to send SMS messages in case of emergency to the saved numbers in the microcontroller. They have analyzed existing electronic aids, listed out their limitations and modified their approach, they also aim to develop emergency trigger alert system along with the design.

Sharma et al. [7] in their paper "Multiple Distance Sensors Based Smart Stick for Visually Impaired People" have also worked on an economically low-cost solution and supposedly durable and accurate smart stick to help the blind navigate both indoor and outdoor unstructured environments. They have considered Indian demographic into context. Their stick consists of two ultra-sonic sensors pointed up and down and have used a novel algorithm to get an idea about distance, height, and location in front of the stick of the objects. It gives feedback to the user through vibration in hand and audio in the ear of the person. The wireless connection between the earphone and the stick was set up using Bluetooth. Different vibration frequencies and different audio tracks alert the person about the obstacle's distance. Different people have conducted real-time experiments in various environments to observe the stick's accuracy and the results are quite encouraging. This paper reports accurate coordination and communication between sensors, motor, controller, Bluetooth modules and other components to build the blind people a smart stick. The developed stick's two main goals were to increase mobility and also, to generate wireless audio output after successfully detecting an obstacle.

Sharma et al. [8] in their paper "Smart Cane: Better Walking Experience for Blind People" have used two types of sensors, the infra-red and ultrasonic sensors. There's a water sensor at the base of the smart cane that detects and dodges puddles. It activates the sound system and the vibration motor when it recognizes any obstacle. The sensor passes this information to the microcontroller when obstructions are detected. This information is then processed by the micro-controller and calculated if the object is close enough. The micro – controller sends a signal to sound a buzzer in case the object is close. In case of an emergency, the GPS system provides information on the current location. After their implementation of this, they have also focused on the product end of the device to manufacture it at a low cost. They have named their startup as "E-Stick" and they describe it as - "E-stick is an electronic interface which has been intended for the visually challenged individuals to make them walk more productively by giving the sign and voice message about the nearness of any obstacles on their way." Agarwal et al. [9] in their paper "Electronic Guidance System for The Visually Impaired -A Framework" have worked on developing an idea of a plausibly implementable product. The main features would be to calculate the distance between obstacle and user including the detection of staircase, a distress signal indicating user needs help using the SOS feature that would incorporate an emergency message. Traffic light Indication and Landmark Detection are the features that would give the surroundings a much broader feeling, enhancing the user's safety and therefore confidence. These potentially affordable systems are expected to reduce dependence on sighted assistance and thus empower the visually challenged.

III. OBJECTIVES

Our Goal here is to create a Personalized assistant to the user, something that they can interact naturally with their voice to mimic the aid of an actual human assistance while they are on the move. A consent-based voice interaction model which would accept and respond to navigation queries from the user and will intelligently inform them about what they should be aware of. The major focus here would be on the reliability of the device, to ensure the maximum fail proofing. To help the visually impaired with basic indoor navigation and outdoor commuting. To provide the user with consent based visual information in auditory format. Fully functional Natural Voice assistance to aid the user with one to one speech-reply feedback system. Intelligent voice assistant that is consent based and smart trigger-based reply. Priority to moving vehicles by adding more weights to the vehicle cluster class. Optimized Detection Efficiency through Single pass NN and detection cycle.

IV. DRAWBACKS OF PREVIOUS MODELS

To keep improvement and advancement in perspective we have focused our research to papers and patents published in the very recent years and which are very specific to a medical aid to help the blind commute. Through extensive research and reviewing all the above techniques mentioned in the papers we have made a list of limitations these systems, though allegedly state of the art, are not able to address problems that our current model is aimed at solving. Most implementations using Ultrasonic sensors and Infra-red sensors for proximity detection through light or sound waves and these don't perform as well as when they are in their ideal scenarios. These are basic sensors which can't be depended upon for a serious task of navigation and commuting outdoors where there is traffic. For instance, ultrasonic sensors don't work so well in loud traffic and start to give wrong triggers which defeats the purpose of safety and is definitely not reliable to walk near traffic. Infrared sensors don't work too well in bright sunlight as even sunlight contains a little bit of IR wavelength and this interference causes false triggers when using the device in sunlight. In addition to this these sensors don't have a wide field of view as compared to a normal human eye and can only detect in linear 3D conical spaces, the use of more sensors can help with this but that would require algorithmic effort to address many exceptional cases in order to eliminate any pseudo inputs. Another major area to focus is how the device communicates with the user, most devices mentioned above have a haptic buzzer feedback or an auditory buzzer feedback system which operate on triggers. These provide linear and vague feedback to the user which lacks a lot of information, and most of the times the user is also being annoyed with too much information and the system keeps buzzing all the time confusing the blind user to not think and act. There is no active form of communication with the device which leaves the user with limited information as he/she won't be able to trigger on consent to get information.

IV. METHODOLOGY

Our project is dependent on Two core architectures working together in congruence. We want to emphasize on two main features of the device, its robust Reliability and humane natural interactive voice assistance. The high reliability of the device in recognizing/detecting/predicting object motion provided by the multi-threaded scene building approach to reinforcement networks and the natural interactive voice assistance provided by context-tree based answering algorithm. The object detection CNN is pre-trained on the dataset. Since our model is a reinforcement learning model, it will be more optimized the more it is used in a certain area or with certain types of objects. Since a visually impaired person doesn't travel over long distances and will be localized to a certain neighborhood, the recurrent structure allows the network to adjust the weights of the objects and people to the type of objects and people of that area.

The master thread executes the object detection CNN indefinitely, its time delay is adjusted automatically on how many objects it captures per shot taken. So initially if it starts with 1 shot per 3 seconds, after capturing 3 shots if it sees more entropy then it increases its frame rate. Our voice assistant is interactive, so the user can ask it any question he likes regarding navigation when he is confused. For every question asked, another thread is created which runs the CNN and is fitted in a parallel layer to the previously created tree structure.

This is to take information from the tree and provide accurate results to the user based on what time frame he is in and what it has predicted.

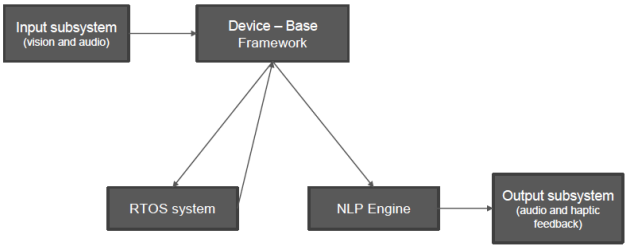

Fig. 1. Abstract Architecture on a high-level schema

The Input Subsystem consists mainly of the camera module to take in visual input of the surrounding and the auditory system to take in audio input of the user. The visual input taken in form of images 1 every 3 seconds is then input into the Base code framework. This is then transferred to the object detection module, which outputs the respective scores of the images and classes detected back into the framework. As we can see in Fig.1 above This information is then processed and analyzed to check if it satisfies the criteria to be classified as an object instance. If true then this information is passed on to the NLP module where it is phrased into a sentence in English. This is then fed onto the output module and then into the speaker or earphone to the user. The audio request given by the user is input into the base framework, which is then analyzed and then depending on satisfying a certain criterion triggers the object detection system.

A. Object Detection Neural Network Architecture

Previous methods of object detection used the help of RCNN – Region based Convolutional Networks. RCNN detects different regional proposals and thus performs multiple predictions in a picture for different regions. The technique we are using is YOLO9000 which works like a FCNN (Fully convolutional neural network) and passes the image of 'n x n' resolution only once through the entire network and the output prediction is achieved. The output is structured in an 'p x p' grid and for each grid - 2 bounding boxes and class probabilities are given. The object detection task is now made into a single regression problem. Multiple bounding boxes and class probabilities for those boxes are simultaneously predicted by a single convolution network. It trains on full images and optimizes the performance of detection directly. This unified model has several advantages over traditional object detection methods. We don't need a complex pipeline because we classify frame detection as a regression problem. In order to predict detections, we simply run our neural network on a new image at test time. The base network runs on a NVIDIA 950 m GPU at 15 frames per second without batch processing. When making predictions, it causes the image globally. Unlike sliding window and region based proposal-based techniques, it sees the entire image during training and testing time, thus encoding contextual information about lasses and their appearance implicitly.

Fast R-CNN, a method of top detection, mistakes background patches for objects in an image because the larger context is not visible. This technique, compared to Fast R-CNN, makes less than half the number of background errors. To predict each bounding box, the network uses features from the entire image. It also simultaneously predicts all bounding boxes for an image across all classes. This means that our network causes the entire image and all the objects in the image globally. The design allows training end-to-end and speeds in real time while maintaining high average accuracy.

The input image is divided into M x M grid in such a way that if the grid cell is on the center of the object then it is responsible for detecting that object. Bounding boxes B and confidence values for those boxes are predicted by each grid cell. These scores of confidences reflect how confident the model is that the box contains an object and how accurate the box is to predict. In formal terms, we define trust as Probability(Object) If there is no object in that cell, the scores of confidence should be zero. Otherwise we want the score of confidence to be the intersection between the predicted box and the ground truth. Each bounding box has five predictions: x, y, w, h, trust. The codes (x, y) are the center of the box in comparison to the grid-cell boundaries. The width and height of the whole image are predicted. The confidence forecast is the IOU between the forecast box and any foundation truth box. Each grid cell also predicts the probabilities of C conditional class, Probability(Class i). These probabilities are determined by an object's grid cell. We only predict one group per grid cell, independently of the number of boxes B. During prediction time the conditional class probabilities are multiplied with the individual bounding box confidence cores – from this for each box, class-specific confidence scores are generated. This contains the information of how well the box fits the object and also the probability of that certain class with respected to the box.

$$Probability(Class\ i/Object) \times Probability(Object) \times IOU) = Probability(Class\ i) \times Object$$

Architecture and Module Framework Our Architecture is divided into 3 parts namely: (i) Mainframe Module (ii) Subsystem 1 (Detection, Recognition) (iii) Subsystem 2 (Voice Assistance)

B. The Mainframe Module

Just like any computer has a CPU, our Mainframe System is a comparable unit. The Mainframe System consists of the Device Script Core which includes the "Pre-trained DCNN Core Module" and the "Decision Making Algorithm Core". Each decision of the Subsystems is taken by the Mainframe Decision Making Algorithm Core. Although the main job of processing the information is done by the Subsystems, the Mainframe System is the one behind the job orders for the same. The Mainframe system is what controls what job needs to be done and by which subsystem.

C. Subsystem 1 – Detection and Recognition

Since our project has 2 major components, so does the Architecture. Each Subsystem carries out a specific part of the job. The job of Subsystem 1 is to detect the objects/ obstructions around the user by using DCCN and the class corpus which is pre-trained into the subsystem. The process in which this is done is – The Camera Module takes continuous images of the surrounding. Each of the image is taken as an input for the mainframe and the subsystem tries to form an output based on certain conditions. The conditions include the location of the objects in the image, the size or description of the object and the distance of the object from the user. For this purpose, let's consider an image with X and Y axis. The X axis is divided into 3 equal parts by the subsystem, these parts determine the location of the object as Left, Center or Right. The position of the object is decided by where the object lies on this X axis. Every object is defined on our system by a Bounding Box. Since we have limited space, our corpus is limited, therefore our project consists of the most extensively used objects in our day-to-day lives. Ex: Table, Chair, Person, Bottle, Mobile phone, Camera, etc. When the camera is confronted by an object, the Subsystem 1 as to decide what object it is, based on the Prediction Class Vector the type of object is decided using the Decision-Making Algorithm Core. When an object is detected, the information is relayed to the Decision-Making Algorithm Core in the Mainframe Unit, which decides what class fits the object aptly. This information is then processed by the Subsystem 1 which calculates the distance of the object from the user i.e. from the camera. Based on how close this object is to the user, the information is passed on to the Subsystem 2.

D. Subsystem 2 - Voice Assistance

When the Subsystem 2 gets information from Subsystem 1 about an object that is immediate to the user, the subsystem passes the data to the Voice Assistance Engine which generates the phrase data. The whole idea of using this is to pass on the information to the user. We need to understand that the user is visually impaired therefore the only way for him to get an idea of the obstacle ahead of him is through a voice assistant, therefore as soon as an obstacle is detected by the user, the kind of obstacle is mentioned to the user in the form of a dialogue, "There is a table on your right" or "There is a person to your left". The role of Subsystem 2 does not end there. Many a times the user wants to know where a certain object is or if there is any object in front of him, in such a case our Voice Assistance System tells the user what is in front of him or the location of any object he is looking for. This has a different methodology itself. In such a case, where the user is asking a question, the question is broken down using Natural Language Processing, and the data is passed on to the Decision Mainframe, which eventually forwards the data to the Subsystem 1. The process of the subsystem 1 is repeated and any object if detected is mentioned to the user. For example, if the user wants to know whether there is anything to his/her right. The user simply asks the Voice Assistance through the click of a button "Is there any object to my right", this sentence is broken down and the important data is used, here the data used would be "object" and "right", this data is relayed to the Mainframe Decision Making Core, which then uses the Subsystem 1 to detect any object on the right side of the most recent image. Based on this, the information is forwarded to the user in the form of speech by the Assistant using the speaker/audio out module.

Apart from the technical goals we have set for the device, we also want to achieve the following which makes our device unique.

1) Understandability and Usability – The device is user friendly and allows even technically challenged people to use it. With personalization, it becomes easier and homelier for the user.

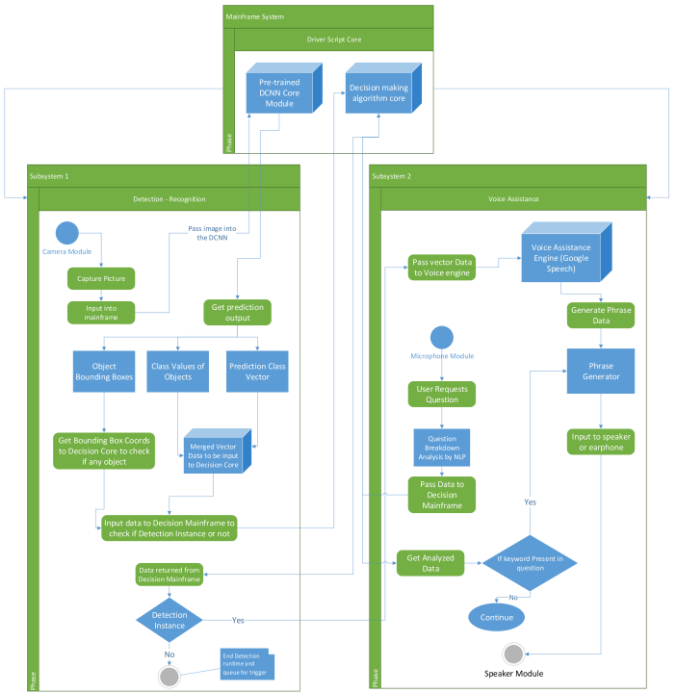

Fig. 2. Complete detailed architecture and Individual Module Framework

2) Performance – The system has a huge data in terms of classes that require higher end hardware, therefore the performance is completely based on the hardware used.
3) Stability and Reliability – The system is such that the false positives and false negatives are minimized, so as to improve precision.
4) Robustness – Each individual unit is quality tested and the entire device goes through rigorous training scenarios for a completely reliable robust performance delivery.
5) Maintenance and Support – Maintenance is required in the form of algorithmic improvements via software updates from time to time.
6) Flexibility and Customizability – The code used is through an open-source platform and the hardware is modular therefore any customizations are possible based on user requirements.
7) Safety and Modularity – The system doesn't contain any user data therefore and is for general use, therefore the safety isn't compromised
8) User-Friendly and Cost- Effective – The entire system is setup in a single package for layman customers to buy.

V. HARDWARE FOR SIMULATION

1) 2.7 Ghz Quad Core Processor A faster 6th Generation Intel i7 CPU for handling the driver core algorithm and also the transfer of data to the memory of driver core and between individual modules. • CPU helps to feed GPU with enough data and read/write files from/to RAM/HDD during training. • If your CPU is weak, it can only feed as few data as possible thus can't keep up with your powerful GPU. • Ideally DL training systems should have CPU with maximum number of processing cores to handle more work to catch up with a GPU.
• CPU is important, for running prerequisites to run deep learning, such as importing data, data cleansing, visualization, statistics, loading data to GPU and so on. • While training deep learning, if the data and intermediates, including weights and calculations in forward and backward propagation, are all loaded to GPU memory, CPU might process very minimum.
2) 16 GB RAM • Memory is today one of the major challenges facing deep neural networks (DNNs). • Researchers are struggling with the DRAM devices ' limited memory bandwidth that today's systems need to use to store huge amounts of weights and activations in DNNs. • The capacity of DRAM also seems to be a limitation. But these challenges are not as they appear to be. • With processor chips specialized for serial processing and DRAMs optimized for high density memory, computer architectures have been developed. • The interface between these two devices is a major bottleneck that introduces limitations on latency and bandwidth and adds significant power consumption overhead.
3) 200 GB Hard Disk Minimum amount required to store the dataset, pretrained models and the code repository.
4 GB VRAM Graphics Card for NVIDIA CUDA Deep Learning • GPUs have additional advantages over CPUs, these include having more computational units and having a higher bandwidth to retrieve from memory. • Furthermore, in applications requiring image processing (i.e. Convolution Neural Networks) GPU graphics specific capabilities can be exploited to further speed up calculations. • GPUs work well with NN computations because (1) GPUs have many more resources and faster bandwidth to memory (2) NN computations fit well with GPU architecture. • Computational speed is extremely important because training of Neural Networks can range from days to weeks. • Many of the successes of Deep Learning may have not been discovered if it were not for the availability of GPUs.
4) 30+ FPS Medium Speed Camera Module for RPi. This is the main Input unit, the Camera module interfaced with the Raspberry pi through CSI (Camera Serial Interface). The sensor itself has a native resolution of 5 megapixels and has a fixed focus lens onboard. In terms of still images, the camera is capable of 2592 x 1944 pixel static images, and also supports 1080p30, 720p60 and 640x480p60/90 video. The captured camera input either in form of a faster pic per second scheme or a slower dynamic video input stream method.
5) Omni Directional Microphone This is for the user to communicate with the device for consent-based queries which helps the user get more insight and information about his vicinity as his order demands.
6) Loud Speaker - for Voice Assistant Output The speaker or the earphones, connected over wire or Bluetooth, is the only source of output given to the user, hence each sentence it speaks out needs to be carefully altered to make more sense, powered by the

Google NLTK API. Hence the voice outputs are carefully pruned to make it concise at the same time give maximum amount of information to the user.

7) Raspberry pi 3 The Raspberry Pi 3 – Model B quadcore .2GHz 64Bit SoC and onboard WIFI and Bluetooth is a full-fledged linux development board. The Pi 3 has two major upgrades. The first is a next-generation Quad-Core Broadcom BCM2837 64-bit ARMv8 processor. This processor has speeds of up to 1.2GHz compared to the previous 900MHz on the Pi 2. The second upgrade is the addition of a built-in BCM43143 WIFI chip, allowing the Pi 3 to go wireless without additional peripherals. No more WIFI adapters. The Raspberry Pi 3 is also an excellent IoT solution with onboard Bluetooth Low Energy (BLE).
The Pi maintains the core algorithm and also the voice interaction module and it is responsible for the prominent communication between the modules. All of the decision making and the main check for whether to inform the user about the obstacle in his path is all handled by the Pi.

8) GPS Sensor for R-pi3 This is a complete GPS module that is based on the Ublox NEO-6M. This unit uses the latest technology from Ublox to give the best possible positioning information and includes a larger built-in 25 x 25mm active GPS antenna with a UART TTL socket. A battery is also included so that you can obtain a GPS lock faster. The Ublox NEO-6M GPS engine on this board is a quite good one, with the high precision binary output. It has also high sensitivity for indoor applications. UBLOX NEO-6M GPS Module has a battery for power backup and EEPROM for storing configuration settings. The antenna is connected to the module through a ufl cable which allows for flexibility in mounting the GPS such that the antenna will always see the sky for best performance. This makes it powerful to use with cars and other mobile applications.

9) Power Delivery for R-pi The Pi takes input power of 5V and 2.5 Amps and we are using a portable power bank (Battery Pack) to keep the device running. 10) Enclosure for the apparatus. The whole device is fitted into a 3D printed box with lanyards to be worn around the neck. Software for Simulation • Base OS - Linux with Kernel 4.18 - Ubuntu 17.10 • Darknet Open Source NN Framework • Darknet is an open source neural network framework written in C and CUDA. It is fast, easy to install, and supports CPU and GPU computation. • Base Code Repo – Python Image Processing Libraries - PIL, SciPy, OpenCV, matplotlib, scikitlearn • Process Handling Libraries - Compatibility with Linux - subprocess, psutil, etc. • Cython Libraries for linking the NN with Driver Script • Speech handling - Google Speech Recognition • Python NLP handling Library – NLTK • Cloud Services • Google Speech Recognition Servers • Google Text handling servers.

VI. RESULT AND DISCUSSION

A. Testing

1) Objective Testing This is basically a QA (Quality Assurance) Check against the objectives of our project and how the device performs against those objectives. We check based on how reliable the device is in different framed scenarios and random environments, like in light to moderate traffic, stroll on the foot path and indoor navigation. We also do a limit testing, to test under harsh and unpredictable conditions to see how well the device fairs and plot accuracy for analysis and threshold adjustment. By the end of this phase we make sure all the objectives are successfully achieved by our device.

2) Blackbox Testing of the Algorithm Known as Behavioral Testing is a software testing method in which the tester is not familiar with the internal structure / design / implementation of the item being tested. These tests may be either functional or non - functional. In here we give the device to an actual blind or visually impaired user and check its functionality in aiding the user. We make a list of all the objectives that the device has helped the user in completing efficiently and with how much accuracy and ease-of-use. After testing each feature, it then tested for its unified non-OS system, which makes it easy for any individual to use.

3) Result Extraction and Comparison After the full testing, we conducted an experiment with a certain population size in which a placebo device with current state of the art architecture (which uses sensors) is given to half and give our device to the rest of the population. We then made similar scenarios for both the users and mapped out the factors that distinguished our device from the others. For each task to be accomplished by the user, we noted down timeframe and ease-of-use of the respective devices and analyzed how our device is performing against the rest.

B. Tradeoffs and Limitations

The device although more advanced in accuracy and working model than other state of the art products available for aiding the blind, it does have initial limitations due to manual prototyping and non-industrial embedding approaches.

1) It cannot be trained with a greater number of classes due to performance and memory limitations.
2) It needs to be charged enough to work as it cannot work for a full day of commuting – battery life will last for a maximum of 5 hours.
3) Maps information is reliant on google maps and positioning is reliant on the current GPS as there is no additional positioning systems.
4) It is currently not so portable to carry, but can be made portable with custom circuitry and custom embedded sensors.

Fig 3. Final Prototype device, note that the screen is only for visualization of the workflow, the visually impaired user is going to use the earphones and have a normal conversation pertaining to the task sheet.

C. Future Work

Before commercial production, the design is advised to be improved. The following are some improvements that might be made:

1) Increase camera range and frame quality and implement a technology to determine the speed of obstacles approaching.
2) Offline stored maps and additional positioning services for accuracy like GLONASS and compass.
3) Synchronization with various internet-available navigation
   software applications to select new, unprogrammed destinations as well.
4) Advancements on making the device more portable and light weight.
5) Developing a lighter algorithm for lower power consumption and faster working – to decrease the power consumption of the device.
6) Adding directional cameras for panoramic vision analysis, even one on the back to alert about obstacles incoming from the back.
7) Option to add in more languages and choose other languages.

VII. CONCLUSION

With the proposed architecture, if built with the highest degree of precision, the blind will be able to move from one place to another without helping others. If such a system is developed, it will act as a basic platform in the future that will be cost - effective to generate more such devices for the visually impaired. The prototype developed gives good results in detecting distant paced obstacles in front of the user. The solution developed for visually impaired people is a moderate budget navigational aid. But minimizing costs leads to performance compromises. In this project we have worked on developing an advanced solution to aid the visually impaired by creating a Personalized assistant to the user with which they can interact with. It can detect the presence of objects, record and calculate their location, recognize that particular object and if it is in the path of the user, communicate with the user to inform him about the obstacle blocking his path. It can also take in audio requests from the user to manually activate and answer back to the user. It works by using its object detection module with high reliability training accuracy to detect object boundaries in motion per frame and once the boundary box crosses the threshold accuracy, recognizes the object in the box and passes the information to the core of the system where it verifies whether or not the information needs to be passed on to the user, if it passes the converted speech The voice interaction model is consent-based, accepting and responding to user's navigation queries and smartly informing them of the obstacle that needs to be avoided.

ACKNOWLEDGMENT

We would like to thank Dr. Swarnalatha Purushotham and VIT University for her guidance and constant push for the analytic approach towards the motive of this project. She has helped us throughout till the completion, giving us feedback and suggestions for improvements and has also managed the publishing end of our paper. The journals from IITs in India regarding our similar research topics have provided great insights into how to make our approach better, we are thankful for the research front of Indian Institutes. We would also like to thank VIT University for giving us this wonderful opportunity to think creatively and apply the knowledge that we gain through the courseware.

REFERENCES

1. Ray, Kumar S., Sayandip Dutta, and Anit Chakraborty. ”Detection, Recognition and Tracking of Moving Objects from Real-time Video via SP Theory of Intelligence and Species Inspired PSO.” arXiv preprint arXiv:1704.07312 (2017).
2. Anwar, Ashraf, and Sultan Aljahdali. ”A smart stick for assisting blind people.” IOSR Journal of Computer Engineering 19.3 (2017): 86-90.
3. Motwani, Tanvi S., and Raymond J. Mooney. ”Improving Video Activity Recognition using Object Recognition and Text Mining.” ECAI. Vol. 1.2012.
4. Kocaleva, Mirjana, et al. ”Pattern Recognition and Natural Language Processing: State of the Art.” TEM Journal 5.2 (2016): 236-240.
5. Yang, Yezhou, et al. ”Robots with language: Multi-label visual recognition using NLP.” Robotics and Automation (ICRA), 2013 IEEE International Conference on. IEEE, 2013.
6. Agarwal, Ankit, Deepak Kumar, and Abhishek Bhardwaj. ”Ultrasonic stick for blind.” International journal of engineering and computer science 4.4 (2015): 11375-11378.
7. Sharma, Sharang, et al. ”Multiple distance sensors based smart stick for visually impaired people.” 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2017.
8. Sharma, Tushar, et al. ”Smart Cane: Better Walking Experience for Blind People.” 2017 3rd International Conference on Computational Intelligence and Networks (CINE). IEEE, 2017.
9. Agarwal, Namita, et al. ”Electronic guidance system for the visually impaired—A framework.” 2015 International Conference on Technologies for Sustainable Development (ICTSD). IEEE, 2015.
10. Jiang, Rui, and Qian Lin Li. ”Let Blind People See: Real-Time Visual Recognition with Results Converted to 3D Audio.” Proc. International Conference on Computer Vision. 2015.
11. Redmon, Joseph, and Ali Farhadi. ”YOLO9000: better, faster, stronger.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

AUTHORS PROFILE

Alisetti Sai Nikhil is a B.tech undergraduate of Computer Science and Engineernig at VIT University, Vellore. Apart from excellent academics he maintains a higher ground in terms on international research working at National Univesity of Singapore and Shantou University, China. His main research areas are General Purpose AI, Deep Learning, Computer Vision, Edge AI mobile platforms and cloud computing. He has phenomenal research experience as he worked with PhD and Post Docs on vision processing and medical imaging in his research internship in China and has successfully published a PLOS One article with a 2.8 impact factor. He also has a patent on his name for making a new algorithm for optimal object detection and recognition on edge AI platforms which he used to develop a device to aid the blind commute indoors/outdoors and locate objects indoors. He also made a hardcoded personal assistant and made the entire device run local inferencing and processing without any dependence on the internet or ad-hoc network. He has profound knowledge in applied mathematics and vision processing and is an astute in deep learning. He has also recently won the G.D. Naidu Young Scientist award issued for his research project evaluated by IIT Bombay and industry experts.

Dr. Swarnalatha Purushotham completed her M.Tech. Degree in 2006 with first rank in Computer Science and Engineering at VIT University, Vellore. She pursued her doctoral programme in the same institution and obtained her Ph.D. Starting as a Teaching Assistant in 2001, she had her stronghold in Image Processing, Artificial Intelligence and Software Engineering and was awarded the Research Award five times at VIT University, Vellore. Her work on Image Processing for medical and satellite images bought her a funded project on "Development of approaches for geometric, radiometric and atmospheric correction of remote sensing data projects for multi-date data analysis" by Space Application Centre-Ahmedabad. She immersed herself in conducting research in chosen area of specialization, guiding Ph.D. projects and M.Tech. students and teaching at undergraduate and post graduate level. Over 70 B.Tech students, 30 M.Tech students, and 100 MS students stand testimony for her productivity in Image Processing, Artificial Intelligence and Software Engineering research. She published more than 75+ papers in national and international journals and conferences. She is a Member of many Professional Societies such as Life Member of Computer society of India (CSI) Professional Member of Association of Computing Machinery (ACM), Senior Member of IEEE and Member of ACEEE.

Lav Mahtani is an undergraduate student of Computer Science major at VIT University, Vellore. Managing a whole company by his hand right after graduating from VIT, Lav has always had a keen interest in business, the market and outsourcing. He has a stronghold in business analytics and IT sector industry trend analysis and has been managing his company for about 4 years now. He has a knack for learning about the industry trend and how the market value of products varies and how to achieve return on investment. He also has a deep knowledge of embedded systems hardware and has contributed in the development of the hardware assembly and functioning of the model.