International Journal of Scientific Research & Engineering Trends
Volume 5, Issue 6, Nov-Dec-2019, ISSN (Online): 2395-566X

Eye Can Help - An Assistive Eye Wearable for the Blind Designed Using Real-Time Object Detection

Hrituja Khatavkar
Dept. of Computer Science
Symbiosis Institute of Technology
Pune, MH, India

Abstract **- It is a big challenge for the visually impaired to access the surrounding with ease. Given paper presents a tool, an assistive device for the blind that will aid them navigate the surroundings. This device uses real-time object detection to identify the objects in nearby environment and then the is processed to convert to speech to make the understand.**

Keywords- **uses real-time, object detection.**

I. INTRODUCTION

For the visually challenged, independent motility is one of the crucial issues. As indicated by WHO study in 2010 , approximately of 285 million individuals live with visual hindrance everywhere throughout the world, out of which 39 million are visually impaired and 246 million have moderate to extreme visual weakness[1].

It incorporates individuals of the considerable number of ages. In India, as per the investigation of 2015, 8.8 million individuals are visually impaired. It represents right around a fourth of the all out world's visually impaired populace and another 47.7 individuals are outwardly disabled. India remains at the second position in the study of visually impaired populace. Nations like China, where as indicated by the WHO information, the visually impaired populace percent is 18% of the world .[2]The individuals who are totally visually impaired have a feeling of loss of certainty while driving and navigating around. Notwithstanding for the halfway visually impaired individuals, making sense of a specific item at long separations is a difficult task.

Although some alternative approaches have been developed for them to perform the daily routine activities, yet they suffer from problems like navigating from place to place with ease as well as social awkwardness. For instance, it is an arduous task for them to find a particular room in an unknown environment. Moreover, blind and visually impaired people find it tough to know whether a person is talking to them or someone else during a conversation.

Technologies in the field of Computer vision, especially the deep convolution neural network (CNN), have been rapidly developed at an exponential rate in recent years. Use of the state-of-art computer vision techniques to help people with vision loss, proves out to be quite promising. In this project, we intend to explore the possibility of using the hearing sense to understand visual objects. The sense of sight and hearing sense share a striking similarity: both visual object and audio sound can be spatially localized. It is not often realized by many people that we are capable at identifying the spatial location of a sound source just by hearing it with two ears.

To overcome the hardships faced by the visually impaired population of our world , we come up with a concept that would predict the object that is in front of the device. Though ,there are many canes and devices introduced in the past for the blind , our device differs from the rest ,as this device is a wearable and detects an object and converts it into an audio information for the convenience of the user.

Our system composes of several modules. Video is captured with a portable camera device (Ricoh Theta S, Microsoft Kinect, or GoPro) on the client side, and is streamed to the server for real-time image recognition with existing object detection models . The location of the objects is then calculated from the location and the size of the bounding boxes from the detection algorithm. Then, the image that is recognized is converted to text . This text is then converted to sound using a 3D sound generation application based on Unity game engine which renders the binaural sound with locations encoded. The sound is transmitted to the user with wireless earphones. Sound is played at an interval of a few seconds, or when the recognized object differs from previous one, whichever earliest.

II. PRIOR STATE OF ART

This problem of providing some kind of aid for the visually impaired has been looked into since a long time. Hence people have developed various canes, waist wearable's, eye wearable's etc. Ox Sight are the smart glasses that are intended to help sight impaired to

© 2019 IJSRET

navigate independently, avoid collisions and see in the dark. The Ox Sight smart glasses use a unique camera system and computer vision algorithms to detect and highlight objects, separating them from the background in real-time. This allows people with even the minimum amount of sight make good use of their vision in everyday scenarios. It is easy to wear the Ox Sight Smart Specs.

It is as straightforward as putting on reading glasses The present versions of Ox Sight's smart glasses enable users to zoom in, change contrasts and furthermore pick between a basic white feature of articles and obstructions, through to upgraded edges and complexity, a carbonized channel that can improve face identification lastly a splendid and bright picture. All channels give live video: as the client moves their head around, the pictures move as well. The pictures are shown on a transparent lens, enabling the wearer to utilize their vision as typical while as yet observing an improved overlay.. Other people can still see the wearer's eyes, which helps in social situations [3].

YGA leaders have designed the world's most revolutionary smart cane, We WALK. The We WALK consists mainly of an electronic handle, with a regular "analog" white cane inserted into the bottom. They have created a 'smart cane' which attaches to the traditional white cane. It inculcates a number of features into the smart cane. Some of these include an ultrasonic sensor which can detect obstacles above waist and gives some vibrations to warn users of obstacles, as well as detecting obstructions below, in the same way a standard cane does. When coupled with the We WALK mobile app via Bluetooth, the user can use apps with We WALK's touchpad voice menu without holding their phone. While the tip of that conventional cane is used to detect obstacles at the ground level, head-level obstructions get pinged by ultrasound pulses sent by the We WALK. Whenever such hazards are detected, the device lets the user know by buzzing their hand through either of two (left or right) vibration buttons.[4].

IIT-Delhi Creates Affordable 'Smart' Cane for the Blind, Costing only Rs 3,000, the device sends out ultrasonic waves that enable user to detect obstacles without having to touch or run into them. The device detects obstacles at a level above the knee, as opposed to a normal stick which only works for obstacles at ground level. A light weight, detachable unit comprising of an ultrasonic ranger and vibrator was developed which offers an increased range of 3m and detects obstacles above knee level. Distance information is conveyed to the user through vibratory patterns that vary incrementally with changing obstacle distance. The projected cost of the device is under 35 USD making it affordable for users in developing countries. [5]

White cane navigational device for the visually impaired. An improved white cane includes an elongate body having a proximal end adapted to be held by or secured to a human user and a distal tip, and a noise generator located proximate the distal tip, the noise generator being configured to generate sound within a range of frequencies audible to humans. The user is able to interpret differences in reflected sound as the white cane is moved about the user's surroundings to obtain more information about the surroundings than could be obtained using a normal white cane. Additionally, the information about the surroundings may become available to the user before that information might otherwise have been available to the user using a standard white cane.

Location, orientation, product and color identification system for the blind or visually impaired Barcode scanning and digitizing technology incorporated in the tip of a white cane for the blind, which houses barcode reader, sonar, color sensory and audio production mechanisms. Apparatus detects and announces via audio or vibratory output, obstacles, drop-offs, colors or suitable informational barcode sites placed with a plurality of locations or environments for assisting blind or visually impaired users to navigate.

In one embodiment, the apparatus combines the white cane with a barcode scanner and decoder with laser depth finding capacity in the cane tip, an sonar technology within the cane housing; and a separate hand-held unit having color and standard UPC barcode scanners, both the cane and the hand-held unit being connected to the audio output device by a hard-wire or wireless connection. The audio production unit can be pocket or purse carried and emits verbal information via respective barcodes suitable for either component or announces color when the respective scanner is used; hand-held color/barcode scanner or cane can be used separately or simultaneously.[6]

III. OBJECTIVE

Our motive is to construct a device which would be beneficial for the audio aid of the visually impaired people. Hence following aims would be followed to construct it:

1. Design and develop a eye-wearable device that would be user-friendly
2. Use Artificial Intelligence and Deep Learning Algorithms for object detection and user-object collision avoidance
3. Generation of 3D audio output for the user to comprehend the live object in front.
4. Integrating the wearable with a smart phone so as to obtain the proximal data and navigation data.

© 2019 IJSRET

International Journal of Scientific Research & Engineering Trends
Volume 5, Issue 6, Nov-Dec-2019, ISSN (Online): 2395-566X

IV. METHODOLOGY

1.Object Detection Algorithm

To detect the objects and their arrangement present in the surrounding, we study many different existing detection systems that could classify objects and also it's spacial arrangement in an image. Deformable Parts Model (DPM) [7] makes use of filters that slides detection windows over the entire image.

R-CNN [8] uses region proposal methods. It generates possible bounding boxes in an image and then applies various Conv Nets to classify each box. The results are then processed and output better and finer boxes. The test time is slow, storage is large and also the training of pipeline is bit of a complex process. This makes it difficult to fit into our application. Fast R-CNN [9] max-pools proposed regions and combine the computation of ConvNet for each proposal of an image and gives output features of all regions at once.

Based on Fast R-CNN, Faster R-CNN [10] can insert an area proposal network after the last layer of ConvNet.[11] Both methods improve the accuracy of computation and also speed up the computational time. But, the pipelines of these methods are still relatively hard to optimize and training them is a complex process. Considering the requirement of real-time object detection[12], in this project,You Only Look Once (YOLO) model [12] is used. YOLO provides relatively efficient objective detection with a great speed.

YOLO Model

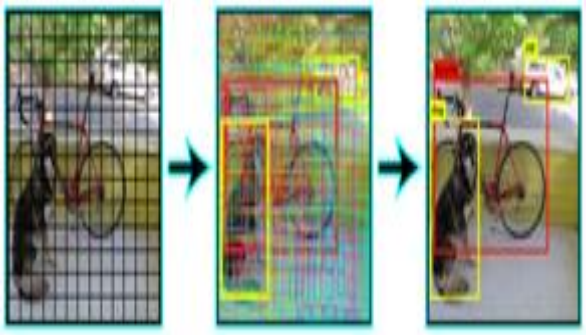

Fig.1 YOLO model divides an image into G×G grid. Each grid cell predicts B.

bounding boxes, and boxes' confidence scores for the prediction and, hence, detect if a class falls in the boxes. The confidence is defined as Pr(object) × IOU predtruth, which represents the confidence of a class in the box and accuracy of the coordinates of the box. Thus, each box has to predict five parameters: x, y, w, h and confidence. Each grid cell also predicts P r(Class i |Object). Thus the confidence for each box is Pr(Classi |Object) × Pr(object) × IOU truth pred =Pr(Class i) × IOU truth. The overall variables to be pre-predicted can be represented as a G× G × (B × 5 + C) tensor. YOLO ConvNet-

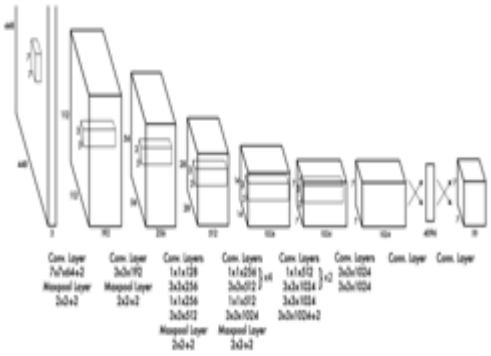

Fig.2 The architecture of Conv Net is shown in Figure.

The network has 24 convolution layers.There are 2 fully connected layers. The ConvNet has to extract features from input images and the fully connected layers then would predict the probability of the coordinates of the boxes and confidence score. The accuracies of the predictions also depend on the architecture of the network. There's also a loss function of the final output. It depends on the x, y, w, h, prediction of classes and overall probabilities. In our project, we use pre-trained YOLO weight to detect objects.

2. Depth Estimation

Once we detect and classify the type of object present in the video frame, we then obtain the distance of the detected object from the user. We use Go Pro Hero3 as it's light-weight, portable and has a large field of view. Also in addition to this, it has high frame rate (about 24 fps) and has wireless compatibility. If the field of view and the bounding box of the object is given, the direction can be then approximated from the central pixel of the bounding box. of the object, the direction can be estimated from the central pixel location of the bounding box.

For the estimated depth or the estimated distance ,we consider a "default" height for any particular class, for example human is assumed to be around 5.5 feet, and car are assume to be 4 feet. We hard code this for each of the 20 classes in our classifier. Then from the height of the bounding box and the default height of the object we can estimate the depth.

3. Data Streaming-

Our project is based on a platform that is capable of processing real-time image. Thus, it is required to have a powerful GPU that could give feedback in no time. The architecture is shown in Figure 5.In this platform, the environment is captured by a portable camera and transfers through HD video link directly to the YOLO model running on a local server machine with high performance GPU. The server detects objects, sends information directly to the unity sound generator and

© 2019 IJSRET

plays the binaural sound. In particular, the environment picture is captured by a portable GoPro Hero3 at 30 frames 1080p resolution. The video is live streamed through the HD video link to the computer server as shown in Figure 6. The HD video link could transfer a high resolution image within 2 miles in 1 millisecond. The object detection engine YOLO then predicts objects in the stream. The YOLO algorithm could process a single image frame at a speed of 4-60 frames/second depending on the image size we send to the engine. The outputs are sent to unity sound generator and the generated sounds are played through wireless ear buds shown in Figure 9. During the implementation, the platform is capable of processing all captured live stream at a minimum speed of 30 frames per second at 1080p resolution.

We need a platform to process, the real-time image. Therefore, we need a powerful GPU that can immediately give a feedback, in no time. For this data streaming, considering the efficiency and computational cost, we use Open CV for doing so. A pipeline is developed that enables us to communicate quickly.

Result Filtering -YOLO gives a yield from the top classes and furthermore gives the probability for every frame. We accept any likelihood above 20% as a certain recognition result. To introduce the outcomes to the user in a sensible way, our algorithm needs to choose whether to speak out a recognized object and at what time. Clearly it's not desirable to continue speaking out the same object to the user regardless of whether the recognition result is right. Similarly, it is also bothersome if two object names are spoken overlapping or intently that the client won't recognize them. To solve the first issue, we assume a cool-down-time of five seconds for each class. For instance, if a person is detected in the first frame and is spoken out, the program will not again speak "person" again until following five seconds.

This is just a sub-optimal solution since it doesn't manage numerous objects of a same class. Practically, if there are two people in a frame, the user must be made aware about the two people, but he shouldn't be informed about the same individual continuously. One potential improvement, which we are as yet working on, is to track the object utilizing overlapping bounding box between the frames. To tackle the subsequent issue, we intend to uphold a delay of half second between any spoken classes. Among every one of the 20 classes of the current YOLO model, we pick the accompanying classes to illuminate the client: "tree", "couch", "person", "pencil", "tvmonitor".

4.3-D Sound Generation-

Fig.3 plug-in for Unity 3D.

We use a plug-in for Unity 3D game called 3DCeption for simulating the 3D sound. We built up a Solidarity based game program "3D Sound Generator" using either a record watcher or TCP socket to get the data about the correct sound clips to be played or also their spatial coordinates. Further, 3DCeption renders the binaural audio effect with the help of the Head-Related Transfer Function (HRTF)[13] in order to simulate the impression of the sound on human body (head, ear, and so forth.) and hindrances, (for example, divider, floor,wall,etc).

5. Testing-

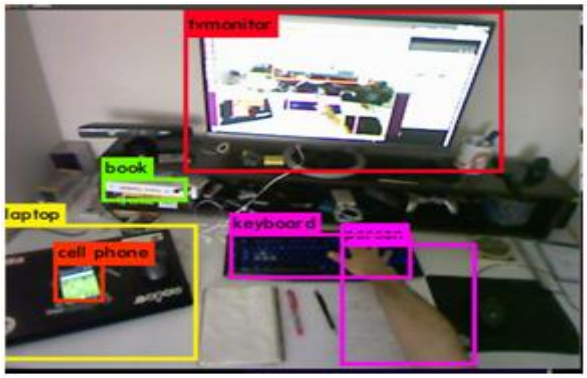

Fig.4 shows the real time object detection performed at by the server.
This is done by using YOLO v1.

Fig.5 camera is used for collecting the data.

Further, when the camera is used for collecting the data YOLO V2 is used. Here it picks up the usual objects like truck, bottle, cup, with a decent accuracy.

© 2019 IJSRET

6. Integrating the Device With Smartphone

Fig.6 Bluetooth the data can be transmitted from the device to the smart phone.

By using Bluetooth the data can be transmitted from the device to the smart phone and vice versa. The Bluetooth Technology is essentially the earliest version of the Internet of Things. The system can then be extended by the use of a smart phone in ways such as incorporating Google maps for navigation and in turn apps like Uber, Ola etc for further advancement.

V. FUTURE SCOPE AND CONCLUSION

There are a lot of equipments available for blind assistance which includes the various versions of white cane and other relevant products. But these hardware's come with their own set of tradeoffs. The cane can be potentially dangerous as it can hit and damage the nearby objects. To provide respite to the blind community, the innovative Eye Can Help eye wear promises to detect objects in the vicinity in the form 3D audio output. Coupled with a smart phone, it reduces the cost and need of extra equipment which reduces the burden on the wearer.

Our work strives to mitigate the problem of independent mobility by blind people and address the issue of obstacle detection and identification. As a part of future work, we intend to incorporate features like Google maps for enhanced navigation This will keep track of places frequented by the user. After Google maps, we can also have other apps like uber(slightly modified) to provide computational facilities. Additionally, certain provisions can be made to signify that the subject is in danger, this can be achieved through some alarm-raising module like emanating loud sounds, so that the people around can sense the danger and rush to the help of the blind subject.

REFERENCES

[1]. Prevalence of Vision Impairment. (n.d.). Retrieved from http://li129 107.members.linode.com/research/statistics-on-vision-impairment/prevalence-of-vision impairment/#worldwide. Who data on Regional visual impairment Retrieved from: https://www.who.int/blindness/GLOBALDATAFINAL forweb.pdf?ua=1

[2]. Yun, E. (2019, November 1). Smart Glasses for the Visually Impaired. Retrieved from https://www.oxsight.co.uk/. Home. (n.d.). Retrieved from https://wewalk.io/

[3]. Papers published. (n.d.). Retrieved from http://assistech.iitd.ernet.in/papers.php#sc1

[4]. Low Vision and the White Cane: A Tool for Fall Prevention. (n.d.). Retrieved from https://www.visionaware.org/info/essential-skills-2/an-introduction-to-orientation-and-mobility-skills/low-vision-and-the-white-cane-a-tool-for-fall-prevention/235

[5]. Felzenszwalb, P. F., Girshick, R. B., Mcallester, D., & Ramanan, D. (2010). Object Detection with Discriminatively Trained Part-Based Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9), 1627–1645. doi: 10.1109/tpami.2009.167

[6]. Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition. doi: 10.1109/cvpr.2014.81

[7]. Girshick, R.B. (2015). Fast R-CNN. 2015 IEEE International Conference on Computer Vision (ICCV), 1440-1448.

[8]. Shao, F., Wang, X., Meng, F., Zhu, J., Wang, D., & Dai, J. (2019). Improved Faster R-CNN Traffic Sign Detection Based on a Second Region of Interest and Highly Possible Regions Proposal Network. Sensors, 19(10), 2288. doi: 10.3390/s19102288

[9]. Redmon, J., & Angelova, A. (2015). Real-time grasp detection using convolutional neural networks. 2015 IEEE International Conference on Robotics and Automation (ICRA). doi: 10.1109/icra.2015.7139361

[10]. J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 779-788. doi: 10.1109/CVPR.2016.91

[11]. Begault DR, Wenzel EM, Anderson MR. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. Journal of the Audio Engineering Society. 2001;49(10):904–916.

© 2019 IJSRET