ALTERNATE VISION ASSISTANCE: FOR THE BLIND

Sarthak Verma[1], Akshita Goel[2]

[1]*Student, Dept. of Information Technology Engineering, Maharaja Agrasen Institute of Technology, IP University, Delhi, India*
[2]*Student, Dept. of Information Technology Engineering, Maharaja Agrasen Institute of Technology, IP University, Delhi, India*

ABSTRACT: *Our paper proposes a system, named Alternate Vision Assistance whose objective is to give blind people the ability to move around in familiar as well as unfamiliar environment, whether indoor or outdoor, through a user friendly voice interface and lead a safe and secure life like a normal being. By this project we not only are able to learn and stay updated with upcoming technologies but also contribute to welfare of the society. Here, a simple user friendly and cost effective smart guidance system for blinds is designed and implemented to guide both blind and visually impaired people in as many possible areas as needed on regular purpose. This device as a prototype model, in different situations is tested, simulating the role of a blind person being who is exposed to new surrounding areas. For example after doing this experiment, a chair that is 3-5 meters away is detected by the user, and then he walked towards it and used it.*

1. INTRODUCTION

Visual impairment and blindness caused by various diseases has been hugely reduced, but there are many people who are at risk of age-related or as congenital visual impairment. Visual information is the basis for most navigational and recognition tasks, so visually impaired people struggle a lot because necessary information about the current environment is not available. Performing or organizing any kind of simple daily activity can be especially difficult; it is not easy for the blind to get aware of surroundings. With the recent advances in inclusive technology it is possible to extend the support given to people with visual impairment during their mobility. For example, in new surroundings, they cannot find a room or a specific area. Or that person cannot judge whether he is able to make an impact to other people in conversation or if other person is interested to listen what he is talking about in a conversation. New computer technologies referring specifically to deep convolutional neural networks here, recently there has been a rapid growth and development. We aimed to use computer vision technologies to help and benefit people with sight problems or losses. Our project proposes a system, named Alternate Vision Assistance whose objective is to give blind people the ability to move around in familiar as well as unfamiliar environment, whether indoor or outdoor, through a user friendly voice interface and lead a safe and secure life like a normal being. As we know that both audio and visual qualities are spatially and coexist. If one sense weakens, the other automatically becomes more trained, trusted and stronger ones. We explore the same possibility to understand surroundings by audio instructions or audio senses. Just by hearing that person can detect the spatial location of objects. Section 3 talked about the generation of 3D sound and different components of our prototype were introduced. The discussions are in Section 4. Then the report is concluded with the Section 5.

2. OBJECTIVE

Key focus is to guide them with an equipment that will be designed to assist them in navigational and detection of the arcade whose uses are not limited to only road guidance but also a supermarket shopper, household, office guide to ensure living at ease using the upcoming technologies like machine learning , deep learning and voice assistance. Visually impaired or blind people cannot identify whether a person is talking to them or paying attention to him or someone else during a conversation.

We aim to create software using machine and deep learning for object detection and depth estimation with a voice assistant to be able to give the corresponding response to the person using it and notify him about the immediate surroundings and provide real time assistance via audio guidance.

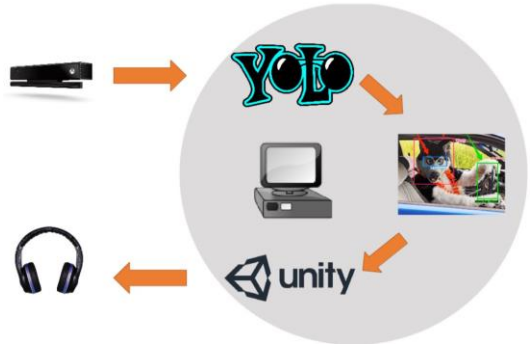

Figure: 2.1 two main focus areas of this project are-

1. Object identification and depth estimation

2. Voice output

So whenever the cane on which mini camera is installed is in use, it can detect the object in front of it and guide in certain ways. The software will make a 3D reality using the object identification and depth estimation algorithms and then notify the person about the objects in front of him and the approximate distance. This is how we provide an alternate vision or simply said a virtual assistance.

3. METHOD

3.1. Vision Simulation algorithm

To efficiently identify surrounding entities, we investigated many existing identification systems which could classify presence of entities and evaluate them at various Sections of an image. Deformable Parts Model (DPM) utilizes root filters which slides detection windows over the entire image. R-CNN utilizes region proposal methods to produce possible bounding boxes in an image. Then, it applies various ConvNets to classify each box whose results are then synthesized again to output finer boxes. It shows a complex training pipeline, prolonged test-time and takes enormous storage space that makes it not very suitable for utilization in this Idea.

Fast R-CNN max-pools proposed regions and unifies the computation of ConvNet for every proposal of the image and produces features of every region instantly. According to Fast R-CNN, it inserts a region proposal network after the last layer of ConvNet. Both methods accelerate the computational time and uplifts accuracy. The complexities for the pipelines of these methods is still high and optimization is very hard taking in account the requirement of real-time identification of entities, for this particular system, we utilize You Only Look Once (YOLO) model. YOLO efficiently provides a comparatively better entity identification with enormously accelerated speeds just by procuring and processing the image in one go.

3.1.1 YOLO Model

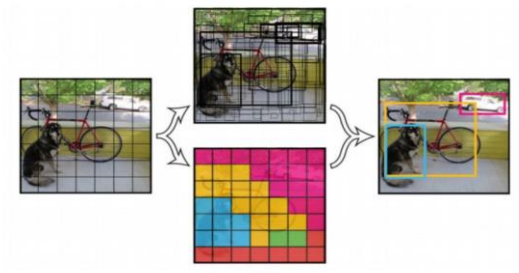

Figure: 3.1 the YOLO Model.

YOLO model partitions a picture into N×N grid unlike previous models which utilize region proposal methods. Each grid cell predicts B bounding boxes and boxes' confidence scores for the prediction and detect if a class falls in the boxes. The confidence is which represents the score and thereafter the possibility of a class of object in that box and accuracy of the box

coordinates. So, every box in all has five distinct parameters to predict the class and confidence score of a particular object.

3.1.2 YOLO Model ConvNet

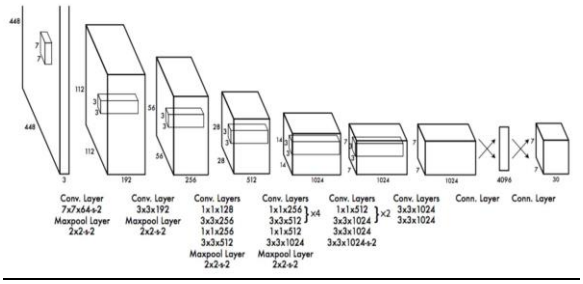

Figure: 3.2 Convolutional Neural Network (CNN)

The ConvNet architecture is shown in Figure 3. The network comprises of 24 convolutional layers with 2 fully connected layers The ConvNet synthesizes the input images and extracts considerable features and the two fully connected layers are employed to predict the probability of the boxes coordinates and confidence score. The architecture of the network hugely uplifts the accuracy of the made predictions. The loss function of the final output depends on the x, y, w, h, prediction of classes and overall probabilities. In our project, we use pre trained YOLO weight to identify entities.

3.2. Depth estimation

In order to combat the difficulty associated with the integration of depth cameras into the pipeline of our project, we go back to the user need for depth information. Firstly, human beings infer direction from binaural sound well, and the relative distance, namely object A being closer than object B or object moving closer and closer between the frames. However, the absolute distance is difficult to devise from binaural sound. This means that our image processing algorithm needs to provide exact directional information and the relative distance, but not the accurate depth.

Thus, we will resort to estimating the direction and relative depth from an RGB image. We will be using GoPro Hero 3 since it's an extremely lightweight carry on camera with a broad field of view, high frame rate and wireless compatibility. Given the field of view of the camera, and the bounding box of the object, the direction can easily be estimated from the central pixel location of the bounding box.

For the approximate depth, we shall assume a 'default' height for a particular class, let's say a human being is assumed to be around 5.5 feet tall and chairs are assumed to be 2.5 feet tall. We will 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 approximate the depth.

3.3. Result filtering

YOLO outputs the top few classes and their respective probability for each frame. We will take any one probability above 20% as a confident detection result i.e. we will set a threshold limit for confidence score for our model in order to filter out the negligible probability objects. To present the results to the user better, our algorithm will also have to decide whether to speak out a detected object or not, and at what time. Obviously, it is undesirable to keep speaking out the same object to the user even if the detection result is correct. It is also undesirable if two object names are spoken in an overlapping manner or so closely, that the user will not be able to distinguish the two.

To solve the initial problem, we will assume a cooldown time of 5 seconds per class. Eg: if a person is detected in the first frame and it is spoken out, the program will not speak out "person" again until five seconds later. This is only a sub optimal solution since it does not deal with multiple objects of the same class. Ideally, if there were two persons in the frame, the user should be informed about both but he does not need to be informed about the same person

continuously. One potential improvement that we are currently working on is to track the object using overlapping bounding boxes between the frames. To solve the next problem, we are planning to enforce a delay of about half a second between any 2 spoken classes.

3.4. Voice Assistance

For voice assistance, we use Google Text to speech API, also known as gTTS API. It is an open source API. We can also make the computer speak a programming language, like Python. Given a string of text, it will speak the written words in English. gTTS is a module and command line utility to that converts spoken text to mp3. It is can be easily be implemented in python 3. The output from the YOLO model and the depth estimation algorithm will be passed to the python code of gTTS code. The result of this code will be a voice telling the blind person about the distance and the object in front of him. Being a google API it can be relied on for best results easily so, it can be directly implemented.

4. DISCUSSION

The prototype we build successfully detect augmented reality and generate 3D audio sound.

Still there are some limitations which are understated with easy expandable solutions.

- PROBLEM: YOLO model can detect objects upto 2-5m distance accurately. If the object is farther it either misrecognized it or the object remain unrecognized. SOLUTION: We can train our model with greater scale ranges of for each object(like a picture of chair from distances 5 meter, 10 meter or even 20 meter). It might be cumbersome for object detection models to classify the object from a picture of an extreme scale, so there should be another approach. We can track the target object when it approaches from a distance.
- PROBLEM: As we are using earbuds or headphones to generate 3D audio output, it blocks the surrounding or ambient sounds making his one of the most powerful sense less of a use. SOLUTION: This can be solved using already invented scientific invention called the conduction earphones which leave the ears open for the surrounding sound.
- PROBLEM: User may be overburdened with the information making it difficult to understand anything about the output heard i.e. the algorithm either generates information about the same object repeatedly after fix interval of time say 5 second or it generates output for different objects simultaneously thereby creating confusion or total havoc. SOLUTION:This can be solved using delayed notification or prioritising things the user is looking for in any specific situation and not focusing much on unrelated classes of objects thereby saving processing and generating desirable outputs only.

5. CONCLUSION AND FUTURE SCOPE

In our project, we have tried to understand the needs of blind or visually impaired people and successfully created a system which is portable and cheap and can be deployed in real life providing assistance to targeted users so that they can lead a more independent and secure life. Using this they can explore the surroundings better . We present a platform using portable hardware system including a camera and installed on the cane and an earpiece and an efficient software which deploy YOLO model which is considered most efficient object detection and classification model by far. The solution provided could take an accurate real time objective detection with live streaming at a speed of 30 frames at an optimal resolution.t. Through our project, we have attempted to demonstrate the potential of using computer vision techniques as a kind of assistive technology.

6. REFERENCES

❑ Rui (Forest) Jiang, Qian Lin and Shuhui Qu's research paper.

❑ Prof. Seema Udgirkar, Shivaji Sarokar, Sujit Gore, Dinesh Kakuste and Suraj Chaskar- research paper

❑ Mathworks for ML and deep learning

❑ LucidCharts for making the flow diagram

❑ Object Recognition by Ming-Hsuan

❑ By Yang University of California at Merced l

❑ Object Detection by Yali Amit and

● Pedro Felzenszwalb, University of Chicago

❑ A Review Paper on Object Detection for Improve the Classification Accuracy and Robustness using different Techniques

● by Divya Patel Research Scholar, Parul Institute of Technology and Pankaj Kumar Gautam Parul Institute of Technology.

❑ Application of Deep Learning in Object Detection

● By Xinyi Zhou, Wei Gong, WenLong Fu, Fengtong Du, Information Engineering School, Communication University of China,CUC Neuroscience and Intelligent Media Institute, Communication University of China

❑ Image Processing and Object Detection

● By Nidhi, NIT Kurukshetra, Haryana, India.

❑ A Review of Detection and Tracking of Object from Image and Video Sequences

● By Mukesh Tiwari and Dr. Rakesh Singhai

❑ Research on Computer Vision-Based Object Detection and Classification

● By Juan Wu , Bo Peng , Zhenxiang Huang , and Jietao Xie

❑ Object Detection Combining Recognition and Segmentation

● By Liming Wang , Jianbo Shi , Gang Song and I-fan Shen

❑ Histograms of Oriented Gradients for Human Detection

● By Navneet Dalal and Bill Triggs X. Object Recognition from Local

❑ Scale-Invariant Features

● By David G. Lowe, Computer Science Department, University of British Columbia

❑ ImageNet Classification with Deep Convolutional Neural Networks

● By Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton

❑ Joseph Redmon and Anelia Angelova, Real- Time Grasp Detection Using Convolutional Neural Networks (ICRA), 2015.

❑ A. Quattoni, and A.Torralba. Recognizing Indoor Scenes. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

❑ Joseph Redmon and Anelia Angelova, Real-Time Grasp Detection Using Convolutional Neural Networks (ICRA), 2015.

❑ A. Quattoni, and A.Torralba. Recognizing Indoor Scenes. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

❑ F. Arman and J. K. Aggarwal. CAD-based vision: object recognition in cluttered range

❑ R. Bergevin, D. Laurendeau and D. Poussart, Registering range views of multipart objects, Computer Vision and Image Understanding

❑ M. Hebert, J. Ponce, T. Boult and A. Gross (Eds.). Object Representation in Computer Vision. Springer-Verlag, Berlin, 1995

❑ M. Hebert, K. Ikeuchi and H. Delingette. A spherical representation for recognition of free- form surfaces. IEEE Trans. Pattern Analysis and Machine Intelligence

❑ H. Hoppe, T. DeRose, T. DuChamp, J. McDonald and W. Stuetzle. Surface reconstruction from unorganized points. Proc. Computer Graphics (SIGGRAPH '92)

❑ A. Johnson, R. Hoffman, J. Osborn, and M. Hebert. A system for semi-automatically modeling of complex environments. Proc. Int'l Conf. on Recent Advance in 3-D Digitalv Imaging and Modeling (3DIM '97), Ottawa, pp. 213-220, May 1997.

❑ A. Johnson and S. Kang. Registration and integration of textured 3-D data. Proc. Int'l Conf. on Recent Advance in 3-D Digital Imaging and Modeling (3DIM '97), Ottawa, 1997.

❑ A. Kalvin and R. Taylor. Superfaces: polyhedral approximation with bounded error.

SPIE Medical Imaging, vol. 2164, 1994.

❑ S. Kang, A. Johnson and R. Szeliski. Extraction of concise and realistic 3-D models from real data. Technical Report CRL 95/7, Digital Equipment Corporation Cambridge Research Lab., October 1995.

❑ V. Koivunen and R. Bajcsy. Spline Representations in 3-D Vision, in Object Representation in Computer Vision, M. Hebert, J.

Ponce, T. Boult and A. Gross, (Eds.) Springer- Verlag, December 1994.

❑ C. Kolb. Rayshade User Guide and Reference Manual. August 1994.

❑ I. S. Kweon, R. Hoffman, and E. Krotkov. Experimental Characterization of the Perceptron Laser Rangefinder. Technical Report CMU-RI-TR-91-1, The Robotics Institute, Carnegie Mellon University, January 1991.

❑ Y. Lamdan and H. Wolfson. Geometric Hashing: a general and efficient model-based recognition scheme. Proc. Second Int'l Conf. Computer Vision (ICCV '88), 1988.

❑ Y. Lamdan and H. Wolfson. On the error analysis of 'Geometric Hashing.' Proc. Computer

❑ Vision and Pattern Recognition 1991 (CVPR '91), pp. 22-27, 1991.

❑ W. Lorensen and H. Cline. Marching Cubes: a high resolution 3D surface construction algorithm. Proc. Computer Graphics (SIGGRAPH '87), 163-169, 1987.

❑ M. Martin and H. Moravec. Robot Evidence Grids. Technical Report CMU-RI-TR-96-06, The Robotics Institute, Carnegie Mellon University, March 1996.

❑ L. Matthies and S. Shafer. Error modeling in stereo navigation. IEEE Jour. Robotics and Automation, vol. RA-3, no. 3, pp. 239-248, June 1987.

❑ C. Montani, R. Scateni and R. Scopigno. A modified look-up table for implicit disambiguation of Marching Cubes. Visual Computer, vol. 10, pp. 353-355, 1994.

❑ H. Murase and S. Nayar. Visual learning and recognition of 3-D objects from appearance. Int'l Jour. Computer Vision, vol. 14, pp. 5-24, 1995.

❑ S. Näher and Christian Urhig. The LEDA User Manual: Version R 3.3. Max-Planck-Institut für Informatik, 1996.

❑ S. Nene and S. Nayar. A simple algorithm for closest point search in high dimensions. Technical Report CUCS-030-95, Columbia University, October 1995.

❑ A. Pentland and S. Sclaroff. Closed-form solutions for physically based shape modeling and recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 13, no. 1991.