Title of Project: Real Time Object Detection and Spatial Audio Generation using YOLO and Head Related Transfer Function

Author: Sukesh Davanthapuram

Advisor: Dr. Jafar Saniie

Co-advisors: Guojun Yang

Xinrui Yu

ABSTRACT

Many visually impaired people stay at home to avoid the challenges and difficulties in navigating from one place to another. Navigating in indoor areas like hospitals, mall is very difficult due to many external factors like noise, crowd, smell etc. Here, we aimed to develop a real time spatial audio and generating software to assist visually impaired people in navigation. YOLO is capable of object detection with good accuracy and helps in identifying different classes of objects and the locations of the detected objects in the frame. Monocular depth estimation helps in calculating the disparity map from a single image. With simple linear interpolation the information about azimuth and elevation angles can be obtained. HRTFs have the capability to convert a mono audio source into a spatial audio using the azimuth and elevation angles. The generated spatial audio's intensity is varied according to the distance of the detected object. The depth of the detected object was calculated with an error in range of 40-60 cm. Overall we can conclude that the real time generated spatial audio was able to clearly specify the object orientation in 2d space and helpful in navigation of visually impaired people.

OBJECTIVE

The main goal of this project was to generate real time spatial audio to assist visually impaired people. The project can be divided into three parts. The first part is object detection. With the help of YOLO all the objects in a given frame can be successfully detected. The second part deals with the disparity map generation and calculation of the azimuth, elevation of the detected object. Finally, the last part focusses on generation of Spatial Audio Using HRTF.

BLOCK DIAGRAM

The proposed solution can be divided into 4 blocks (Figure 1). The frame data from the webcam is sent to YOLO object detection for detecting all the objects in the frame. This frame data is then sent to Monocular depth estimation module which outputs a disparity map. Once the disparity map is generated the depth of each pixel can be calculated. The bounding box data is sent to the azimuth and elevation estimation function which then outputs the required azimuth and elevation index. All this data is sent to the HRTF spatial audio generation module for the spatial audio.

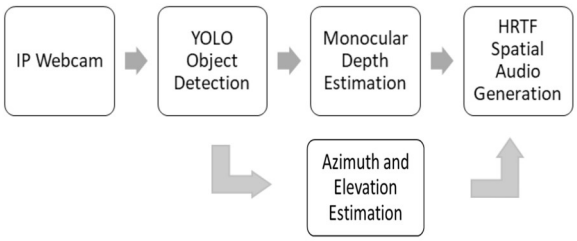


Figure 1. Block Diagram of Project

IP WEBCAM

IP Webcam is an application which converts the smartphone into an internet camera. Once turned on a unique IP address will be generated (Figure 2) and the camera feed is displayed in the website. Using the local IP doesn't require any internet connection. This is a very convenient way to send the camera feed. Almost everyone has a smartphone, with the help of this simple application any individual can access the Spatial audio generation software.

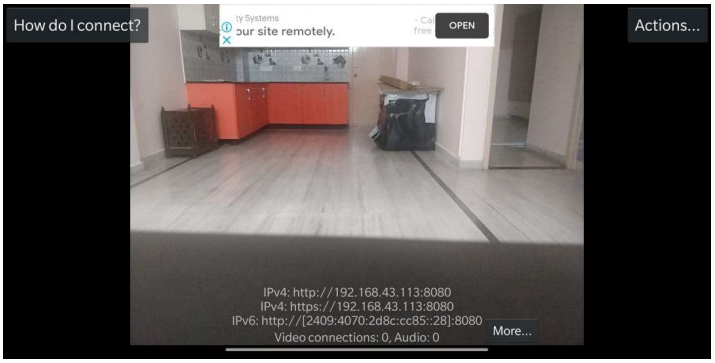


Figure 2. UI Of IP Webcam Application. IP address is shown with white text at the bottom of the picture

As per the above image we can see that the IP Webcam generated both IPv4 and IPv6 addresses. By accessing the mentioned URL, we can get the camera feed.

You Only Look Once (YOLO)

YOLO (You Only Look Once) is capable of detecting objects using convolutional neural networks [1]. There are two tasks involved in object detection. The first is to determine the object location and the second is to classify those objects. There are even other methods to detect objects like the use of R-CNN or its variations. But these are slow and are difficult to optimize. A single neural network is applied to the Full Image. It means that the image is divided into regions and the bounding boxes along with the

probabilities are predicted by the network for each region. Figure 3 shows the result of YOLO and the obtained FPS.

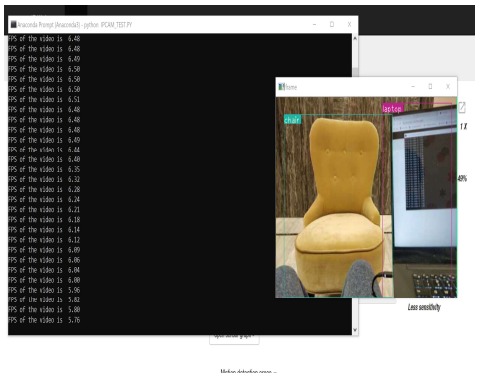


Figure 3. Frame Rate obtained using Laptop's inbuilt Webcam

MONOCULAR DEPTH ESTIMATION

The depth of an image can be calculated to a good accuracy with the help of a stereo camera. This project uses a single camera system. It's a difficult task to get the disparity data from a single image. To tackle this problem Monocular Depth Estimation has been used. Many existing depth estimation algorithms were trained with colour input images and their corresponding depth values [2,3,4]. But in reality, it is no possible to obtain the ground truth depth data of various scenes. Even hardware like laser scanners can be imprecise due the reflections and various other factors. The depth of an image can be estimated with the help of a stereo camera. But the IP Webcam does not have a stereo output. Adding to it many of the smartphones don't have a stereo camera. By using a single image, the depth of the image should be calculated. With the help of monocular depth estimation, the depth of an image can be calculated. Figure 4 shows the real time disparity map of the IP Webcam feed.

Figure 4. Sample Image and Disparity map obtained for IP Webcam Feed (Left to Right)

The model used is trained on KITTI Dataset. The depth of each pixel is calcuated using the formula[5] as below.

D = f*B/d

Where,

D = Depth

f = focal length (in pixels)

B = baseline (in metres)

d = disparity (in pixels)

The baseline for KITTI stereo dataset is 0.54m. The model used for monocular depth estimation has an effective baseline of 0.1 units. So, a scaling of 5.4 was applied for depth prediction.

AZIMUTH AND ELEVATION MEASUREMENT

To measure the azimuth and elevation of a detected object the angle of view of the camera and the resolution are required. With the help of simple linear interpolation, the angles can be estimated.

HEAD RELATED TRANSFER FUNCTION

Humans are capable of spatializing and locating a sound source with a great accuracy. The human auditory system uses various cues to locate a sound source. The time and level differences between both ears will help in spatializing sound source. When a sound travels from the source to the ears various transformations like diffraction, reflections on various surfaces takes place. All these transformations are captured by the Head Related Transfer Functions (HRTF). With these two functions and a mono audio source we can give spatial dimension to a sound.[6]

RESULTS

A small setup containing a bench and a bottle has been arranged. Depending on the position of the bottle with respect to user, a spatial audio will automatically be generated by the software. The Figures 5,6 show the position of the bottle and the spectrum of the generated audio.

Figure 5.1 Bottle is detected to the left of the user and a spatial audio is generated.

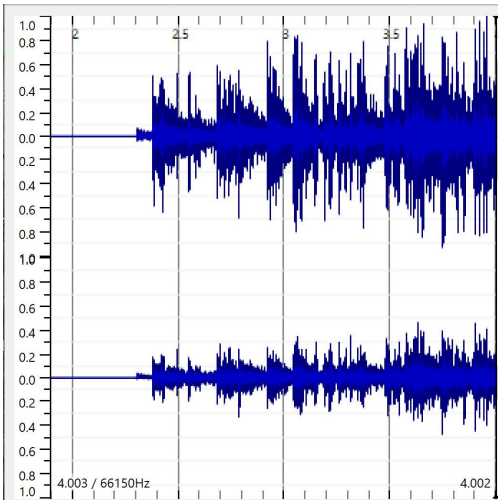


Figure 5.2 Spectrum of Generated Audio

DATA LOGGING

Data logging provides necessary information to make decisions and improve the efficiency of system. The objects detected; its positions are stored in text files.

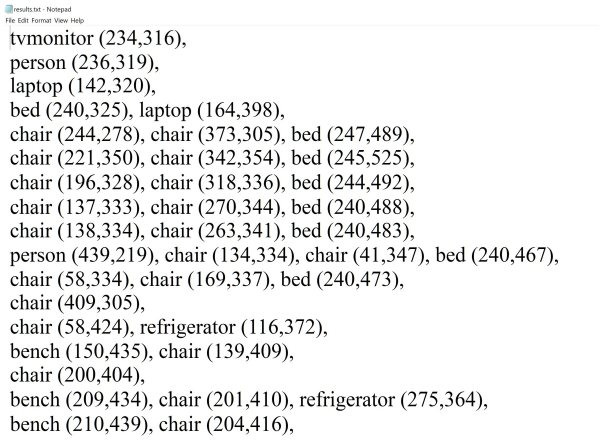


Figure 4.12. Data Logging of the detected Objects

CONCLUSION

The project work has been explained through different sections. The conclusions of various steps involved in this process will be briefly presented.

Rather than giving a low processing device and a webcam to the visually impaired person, it would be better option to choose an IP Webcam. Using this technology is very convenient and many people interested in this technology need not buy any additional hardware.

The YOLO is a powerful object detection model with high accuracy. The bounding boxes formed around the detected object give a very clear idea about the object location. Monocular depth estimation is used to calculate the depth from a single image. This model's disparity map is comparable to that obtained from the stereo camera setup. The depth was calculated with error ranging from 40-60 cm. For large distances this is a very good result.

The azimuth and elevation were calculated using simple linear interpolation and gave good results. With the help of these angles the azimuth and elevation indexes were calculated. The spatial audio was generated using specific azimuth and elevation indexes. Then this audio was attenuated depending on the distance between the detected object and the visually impaired person. With the help of the generated spatial audio the azimuth angles of the detected objects are clearly identified and thus help in navigation of the visually impaired people.

REFERENCES

1. Joseph Redmon, Ali Farhadi, University of Washington, YOLOv3: An Incremental Improvement. https://arxiv.org/pdf/1804.02767.pdf
2. D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In NIPS, 2014
3. L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. In CVPR, 2014
4. F. Liu, C. Shen, G. Lin, and I. Reid. Learning depth from single monocular images using deep convolutional neural fields. PAMI,2015
5. Clement Godard, Oisin Mac Aodha, Gabriel J. Brostow , University College London, Unsupervised Monocular Depth Estimation with Left-Right Consistency http://visual.cs.ucl.ac.uk/pubs/monoDepth/
6. B. Groethe, M. Pecka, D. McAlpine: Mechanisms of Sound Localization in Mammals, https://pubmed.ncbi.nlm.nih.gov/20664077/

We deeply acknowledge with gratitude, the contribution of **BITMAA-NA** and **BIT Mesra** for partly sponsoring our Immersive Summer Research Experience (2017)
at Illinois Institute of Technology, Chicago.

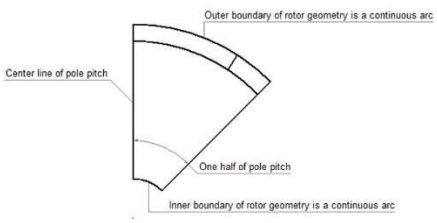
Performance comparison for 2004 and 2010 Prius Traction Motors in a Co-simulation Environment

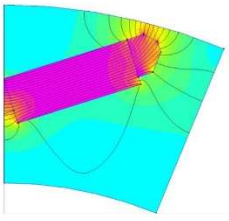
By, Vishesh Sajgotra (BE/10009/17), Electrical and Electronics Engg.

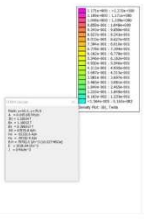
Guide – Dr, Mahesh Krishnamurthy, PhD
Professor and Director | Grainger Power Electronics and Motor Drives Lab | Illinois Institute of Technology | Chicago, IL
tel. +1 (312) 567 7232
kmahesh@iit.edu

Static and Dynamic Performance Analysis of Motors is an integral part in the industrial manufacturing process of Electric Vehicles (EV's). The performance analysis and comparison of previous models helps various EV manufacturers to constantly progress in the field of performance and stability of the electric vehicles. Static performance analysis involves to evaluate the electromagnetic aspects whereas Dynamic performance analysis involves the evaluation of thermal aspects of the motor. The electromagnetic performance to be assessed includes maximum motor torque output for vehicle acceleration and the flux weakening capability for wide operating range under current and voltage limits. Thermal analysis is performed to evaluate the health status of the magnets and windings for the prescribed driving cycles. In the execution part of the project, the report of Oakridge Labs was studied to learn about the motors and methods that were used by them for the analysis and evaluate the performance parameters for the Prius 2004 and 2010 motors respectively. The major constructional differences between both the motors were noted. This process was followed by winding calculation. an approximation was made for the number of poles and number of stator slots to calculate the maximum fundamental winding factor and number of slots per pole per phase using a winding layout program available online

After these pre-requisites, static performance analysis was done using FEMM. For design in FEMM, the DXF file was needed for the Rotor and Stator in order to do the magnetic analysis for the motor. The DXF files were created in SOLIDWORKS 2016 for both the models., Firstly, a single unique segment was designed for each motor component and then the whole CAD was completed by using the circular layout and mirror layout options in SOLIDWORKS and then the DXF file were exported to the FEMM model and the required materials were given to the CAD. This was to be followed by the dynamic analysis but due to network issues and system limitations, it wasn't completed and the project on the basis of static performance analysis.







We deeply acknowledge with gratitude, the contribution of **BITMAA-NA** and **BIT Mesra** for partly sponsoring our Immersive Summer Research Experience (2017)
at Illinois Institute of Technology, Chicago.