International Conference on Innovative Research in Engineering, Management and Sciences
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 4 | Issue 9 | ISSN : 2456-3307

Blind Leap Real-Time Object Recognition with results converted to Audio for Blind People

Dr.Jaya R*1

*1 Department of CSE, New Horizon College of Engineering, Bangalore-560103 Karnataka, India

ABSTRACT

This project tries to change the visual world into the audio world. It has the likelihood to inform blind people about the objects as well as their spatial locations. The objects that are detected at the scene are represented by their names and are then transformed to speech. Their spatial locations are encoded into the 2-channel audio with the help of 3D binaural sound simulation. The system is collected of various modules. The video is captured by a portable camera device (Raspberry Pi with Noir Camera) on the client side. It is then streamed to the server for real-time Object recognition with existing object detection models (YOLO). The 3D location of the objects is determined by the location and the size of the bounding boxes using the detection algorithm. A 3D sound generation application, built on Unity game engine then renders the binaural sound keeping the locations encoded. The transmission of the sound to the user happens with Bluetooth/3.5 jack earphones. The sound is played at an interval of a few seconds or when the recognized object differs from the last one - depends which one is the earliest.

Keywords : Object Recognition, Object detectionYOLO, Raspberry Pi, Unity.

I. INTRODUCTION

Millions of people around us live with impotence of understanding the environment due to visual impairment. Although they develop substitute approach when it comes to dealing with everyday things and routines, they may find it difficult to navigate around and may also be inclined to social awkwardness. For example, it is very difficult for them to look for a specific place or shop in an unknown environment. Blind and visually impaired people may also find it difficult to know whether a person is trying to talk to them or with someone else. Computer vision technologies, especially “deep convolutional neural network”, has developed swiftly over the past few years. It is optimistic to use the state-of-art computer vision techniques to help people with vision loss. In this project, I want to explore the possibilities of using the sense of hearing to understand the visual objects. The sense of sight and hearing share a striking similarity- both visual and audible object can be located spatially. Not many people realize that we are capable of identifying the spatial location of a sound source just by hearing it with our ears. In this project, I have built real-time object detection and position estimation pipeline, aiming at informing the user about the surrounding object and their spatial position using binaural sound.

II. EXISTING SYSTEMS

There exist multiple tools to use computer vision technologies to help assist blind people.

- The "Blindsight" offers a mobile app Text Detective featuring OCR or optical character recognition technology to detect and read text from pictures captured by using the camera.
- The mobile app "TapTapSee" uses computer vision and crowd sourcing in order to define a picture captured by the blind users in about 10 seconds.
- Facebook is also developing image captioning technology to help blind users engage in conversations with other users through pictures.

However, these products were not focusing on magnify general visual sense for the blind people and neither used the spatial sound techniques to further strengthen the user experience. Some work exists in the general scope of sensory substitution.

- Colorblind artist Neil Harbisson developed a device to transform colour information into sound frequencies.
- Daniel Kish, who is totally blind, developed accurate echolocation ability using "mouth clicks" for navigation tasks including biking and hiking independently.
- An extreme attempt of converting visual sense to sound is introduced by the vOICe technology. The vOICe system scans each camera snapshot from left to right, while associating height with pitch and brightness with loudness[1].

However, all these attempts on sensory substitution are reported to be a very difficult learning process. In contrast, I utilize visual recognition algorithms which lead to more direct ways of understanding the objects from a visual scene.The utilization of 3D sound innovation for giving helpful data and helping blind individuals has likewise been examined by researchers. A system was introduced that uses spatial audio to ease the discovery of points of interest in large, unfamiliar indoor environments (e.g. shopping mall). It tries to integrate 3D sound into GPS-based outdoor navigation product. However, no visual recognition has been used in the current done works. The use of object detection techniques can open up new feasibility in ease of indoor navigation for blind and visually impaired people.

III. APPROACH

A. Object detection algorithm

To successfully detect the objects in the surroundings, I investigate several existing detection systems that could classify objects and evaluate it from various locations in an image. Deformable Parts Model (DPM) uses root filters that slides detection windows over the entire image. RCNN uses region proposal methods to generate possible bounding boxes in an image. It then applies various ConvNets to classify each box. The outcomes are then post processed and output finer boxes. The slow test-time, complex training pipeline and the large storage does not fit into the application. Fast R-CNN[2] max-pools proposed regions and combine the computation of ConvNet for each proposal of an image and outputs features of all regions at once. Based on Fast R-CNN, Faster R-CNN[3] inserts a region proposal network after the last layer of ConvNet The two method accelerate the computational time and improve the precision.The pipelines of these techniques are still generally unpredictable and difficult to enhance. Considering the requirement of real-time objective detection, in this project, I use You Only Look Once (YOLO)[4] model. YOLO could efficiently provide relatively good objective detection with extremely fast speed.

B. YOLO Model

Instead of using region proposal method, the YOLO model divides an image into S×S grid. Wherein, 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 defined as Pr(classi | object)× IOUtp , which represents the confidence of a class in the box and accuracy of the box coordinates. Therefore, all the boxeshave5 parameters that can predict: 'x', 'y', 'w', 'h' and 'confidence'. Each grid cell also predicts Pr(classi |Object). Thus the confidence for each box is

Pr(classi |Object) × Pr(object) × IOUtp = Pr(classi)×IOUtp . The overall variables to be predicted can be represented as a S×S×(B×5+C) tensor.

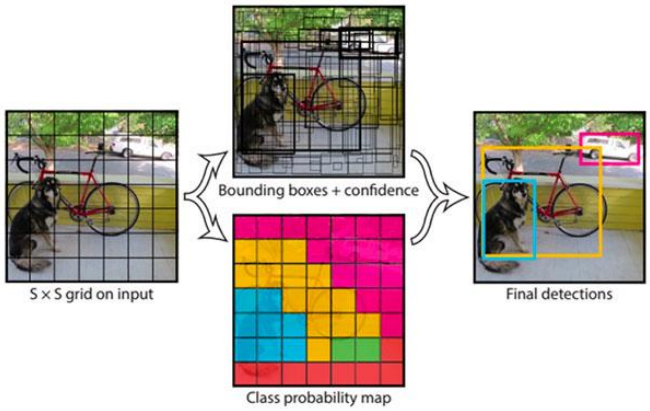

Fig. 1: The YOLO Model

C. Spatial distance

After identifying the type of objects in a video frame, the next step is to obtain the depth or distance of the detected object from the user. First of all, human is good at knowing direction from binaural sound, and the relative distance, namely object A is closer than object B or object is moving closer and closer between frames. However, knowing absolute distance is difficult to deduce from binaural sound. This means that image processing algorithm needs to provide the accurate directional information and the relative distance, but not the exact depth. So,I make use of this to approximate the direction and relative depth from an RGB image. Therefore, giving the camera'sfield of view. The bounding box of the object is also given which helps in finding the direction which can be estimated from the central pixel location of the bounding box. For the estimated depth, I assume a "default" height for any particular class, for example human is assumed to be around 5.5 feet, and chairs are assumed to be 2.5 feet. I 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 I can estimate the depth.

D. Flow of Data

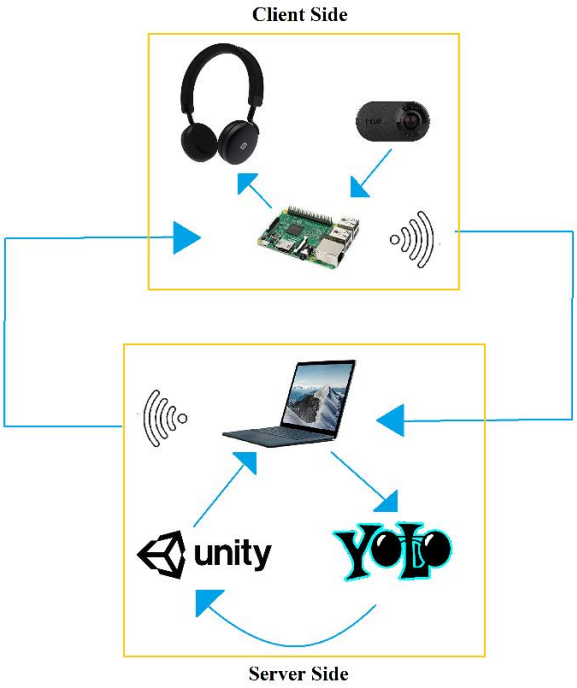

Fig. 2 Data flow pipeline of the system

This project is based on a platform that is capable of processing real-time image. Therefore, it is recommended to use a system to have a powerful GPU which will eventually give feedback in a very short time. A pipeline is developed that enables us to communicate quickly. As Figure 4 shows, a program in local machine extracts raw image from a camera,encodes it into a string and sends through a client to a server. The server fetches the encoded string and decodes it on which trained object detection modalis used to return detected items. The server then sends that information back to the client, which triggers the Unity-based stereo generator to play the 3D sound. The portable camera transfers the HD video 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 plays the binaural sound.

E. 3D Sound Generation

I used a plug-in for Unity 3D game engine called 3DCeption[5] to simulate the 3D sound. A Unity-based game program "3D Sound Generator" is

developed using either a file watcher or TCP socket to receive the information about the correct sound clips to be played as well as their spatial coordinates. The, 3DCeption is used to render the binaural sound effect with the help of the Head-Related Transfer Function (HRTF) to simulate the reflection of the sound on human body (head, ear, etc.) and obstacles (such as the wall and the floor). Since most of the people who have sight may not be aware of the sound localization capability, that is why the reader is recommended to experience the 3D binaural sound effect demonstrations.

IV. CONCLUSION

In this project, I have investigated the need for the blind and visually impaired people. Based on the impetus of the CNN, I developed a blind visualization system that helps blind people explore the surrounding environment in a better way. Easily carried and real time solution is provided in the project. A platform that utilizes portable cameras, fast HD video link and powerful server to generate 3D sounds is also provided. By utilizing YOLO algorithm and advanced wireless transmitter, the arrangement couldperform accurate real time objective detection with live stream at a speed of 30 frames, 1080P resolution.The project provides a vision for hearing. Through this project, I hope to demonstrate the possibility of using computer vision techniques as a type of assistive technology.

V. REFERENCES

[1] David Brown, Tom Macpherson, and Jamie Ward,seeing with sound? exploring different characteristics of a visual-to-auditory sensory substitution device. Perception, 40(9):1120–1135, 2011.

[2] Ross Girshick. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, pages 1440–1448, 2015.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems, pages 91–99, 2015.

[4] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. arXiv preprint arXiv:1506.02640, 2015.

[5] Unity3D:https://docs.unity3d.com/560/Documentation/Manual/AudioSpatializerSDK.html