Synchronous Object Detection with Voice Feedback

Naman Mishra[1], Deepankar Singh[2]

[1,2] UG, Computer Science & Engineering, Babu Banarasi Das Northern India Institute of Technology, Lucknow, Uttar Pradesh, India.

Abstract - *Object Detection is a field of Computer Vision that detects instances of semantic objects in images or videos (by creating bounding boxes around them in our case). In this project, we will convert image to text and then text to speech for the visually impaired person who deserve to live independently by using You Only Look Once V3 (YOLO v3) algorithm that runs through a variation of an extremely complex Convolutional Neural Network architecture called the Darknet with OpenCV and Google Text to Speech, We can then convert the annotated text into audio responses and give the location of the objects in the camera's view. The system will continuously capture multiple frames using a camera on raspberry pi and the frames then converted to audio segment, the obtained results manage to achieve the success of the proposed prototype in giving visually impaired users the capability to understand unfamiliar surroundings, through a user friendly device with this profound object identification model.*

Key Words: *Object detection, YOLO, Deep neural network, OpenCV, Python, Raspberry Pi3, Google Text To Speech.*

INTRODUCTION

A large number of individuals live in this world with the inadequacies of understanding nature because of visual weakness. In spite of the fact that they can create elective ways to deal and manage day by day schedules, they experience certain route troubles as well as social clumsiness. For example, it is hard for them to locate a specific room in a new situation. Furthermore, dazzle and outwardly debilitated individuals think that it's hard to tell whether an individual is conversing with them or another person during a discussion.

Computer vision technologies, particularly the Deep Convolutional Neural Network, were developed rapidly in recent years. Use of the state-of-the-art computer is promising Vision techniques to help vision loss sufferers. In this project we will use the power of the deep neural network and computer vision to give an opportunity to an individual who is visually impaired to see this world. In recent years computer vision technologies have been developed which are very accurate and give promising results [1] presented real time object detection system using a CNN in order to recognize objects.[2]Presented an scanning system based on optical sensors as a talking stick for objects detection in front of the visually impaired. A portable device using a cell phone for visually impaired people is introduced in[3]. In this project we will use the sense of hearing in order to visualize the object kept before the person and the camera. We will use the state of the art "You Only Look Once: Unified, Real-Time Object Detection" algorithm[7] trained on the COCO[5] dataset to identify the object present before the person. Then the label of the object is identified and then converted into audio by using Google Text to Speech (gTTS) which will be the anticipated output. The person will use the output of our system in order to make him aware about his surroundings.

In our project, we will build a real-time object detection with voice feedback system with the goal of telling the user what all is in the surroundings with its spatial position.

1.1 Yolo v3

In this project we have used YOLO v3[8] which is faster than the prior version. It works three times faster, at 320 × 320 YOLOv3 runs 22ms at 28.2 map. It has a similar performance but 3.8× faster. The most notable characteristic of v3 is that it makes 3 distinct scales of detections. YOLO v3 is a fully convolutional neural network and it generates its resultant output by applying a 1 x 1 kernel to a feature map. In YOLO v3, the recognition is obtained by implementing 1 x 1 detection kernels to three-size feature maps at three different regions in the network.

Within each boundary the network predicts 4 coordinates tx,ty,tw,th. Whereas if a cell is offset in the upper left corner of the image by (cx,cy) and prior bounding boxes has pw, ph width and height respectively then the prediction is done [9].

$$bx = \sigma(tx) + cx$$

$$by = \sigma(ty) + cy$$

$$bw = pwe\ tw$$

$$bh = phe\ th$$

1.2 OpenCV

Techniques for Object Recognition in Images and Multi-Object Detection[6] and segmentation is the most significant and testing central undertaking of Computer vision. It is a basic part in numerous applications, for example, image search, scene understanding, and so far. However it is as yet an open issue because of the assortment and multifaceted nature of item classes and foundations.

The most effortless approach to identify and fragment an item from a picture is the shading based techniques. The item and the foundation ought to have a critical shading distinction so as to effectively portion objects utilizing shading based strategies.1

OpenCV[4] usually captures images and videos in 8-bit, unsigned integer, BGR format. In other words, captured images can be considered as 3 matrices; BLUE, GREEN and RED (hence the name BGR) with integer values ranging from 0 to 255.In genuine pictures, these pixels are little to such an extent that the natural eye can't separate.

Typically, one can feel that BGR shading space is progressively appropriate for shading based division. Be that as it may, HSV shading space is the most appropriate shading space for shading based picture division. Thus, in the above application, I have changed over the shading space of the unique picture of the video from BGR to HSV picture.[10]

HSV shading space comprises 3 matrices, HUE, SATURATION and VALUE. In OpenCV, esteem ranges for HUE, SATURATION and VALUE are separately 0-179, 0-255 and 0-255. HUE speaks to the color, SATURATION speaks to the sum to which that individual shading is blended in with white and VALUE speaks to the sum to which that individual shading is blended in with dark.

1.3 Operating System

We are running our project on a raspberry pi 3b+ 1GB LPDDR2 SDRAM variant. The 1GB ram is deficient to run the YOLO v3 algorithm properly with dnn because the operating system is not optimized according to our needs. So, we used a command line OS called Tiny Core in our project. The main benefit of using Command Line OS is that it basically helps us to reserve RAM only for the processing of our project. We installed all the necessary dependencies required to run our project.

1.4 Hardware

We are using Raspberry Pi 3 Model B+ which has Broadcom BCM2837B0, Cortex-A53 (ARMv8) 64-bit SoC @ 1.4GHz and 1GB LPDDR2 SDRAM with class 10 micro SD card where the OS and project is stored, the reason behind we are using the class 10 memory card is that it helps to retrieve data at higher speed so that the project can take lesser time to execute The camera used in our system is Raspberry Pi Camera Module v2 which has Sony IMX219 8-megapixel sensor to feed images at 30 frame per second to this trained model,

1.5 gTTS (Google Text to Speech)

gTTS, a Python library and CLI tool to interface with Google Translates text-to speech API. Converts from document to a spoken mp3 information, a record like article (bytestring) for additional sound control, or stdout. It highlights adaptable pre-preparing and tokenizing, just as programmed recovery of bolstered dialects.

2. Working

We are using Python3 for this project, the camera is initialized by using OpenCV library and the camera starts capturing frames with the rate of 30 frames per second to the algorithm. Then the system uses YOLO v3 which is trained on the COCO dataset and Dark Neural Network (DNN) to identify the object kept before the user.

The object identified is later converted to an audio segment using gTTs which is a python library. The audio segment is the output of our system that gives the spatial location and name of the object to the person. Now by using this information the person can have a visualization of the objects around him. The proposed system will even protect the person from colliding to the objects around will secure him from injuries.

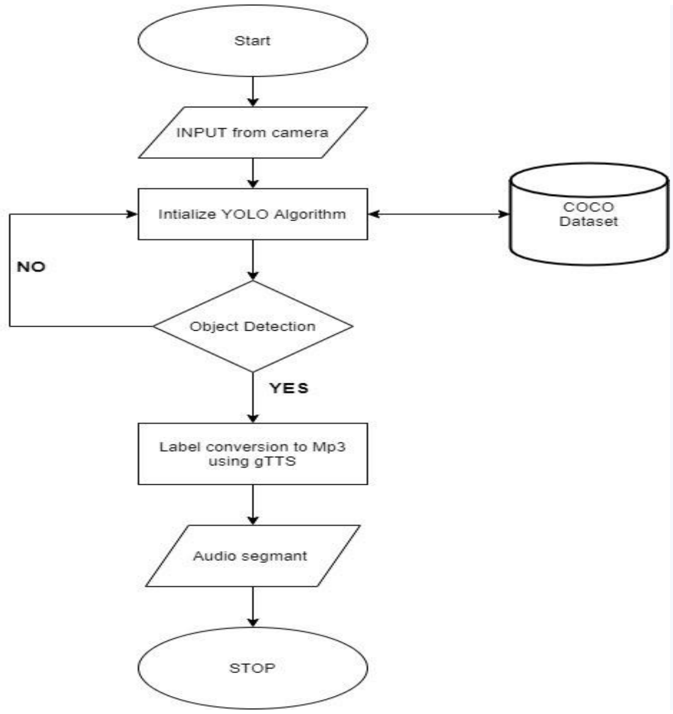


Figure 1: Flowchart of the proposed system

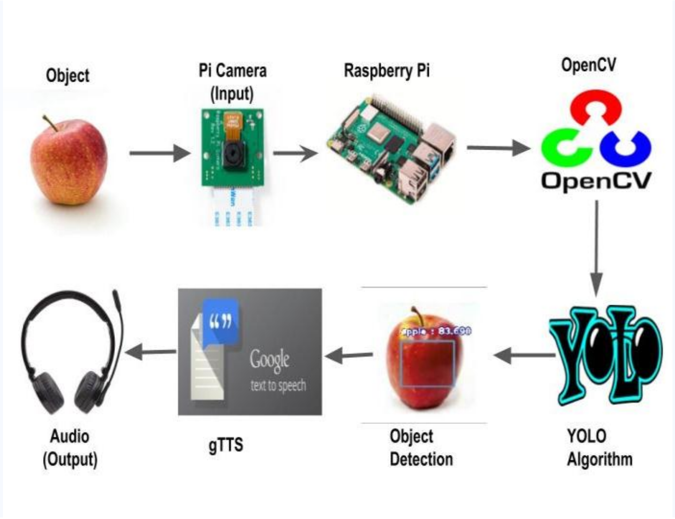


Figure 2: Proposed System

Result and Experiments

The proposed system will be able to identify the object in front of the camera and will later on convert it into mp3 using gTTS. The proposed system is very low cost, FIG 3 shows the whole system which is a Raspberry pi 3b+, Bluetooth headphones and a power bank in order to provide power to the raspberry pi. The hardware implementation is shown in the given figure.

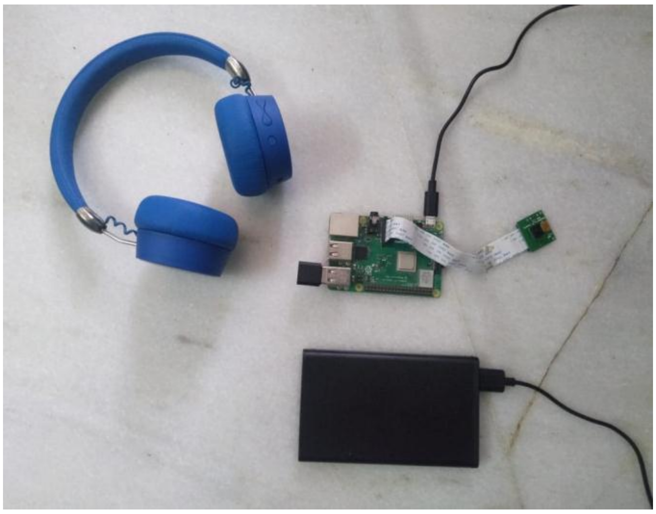
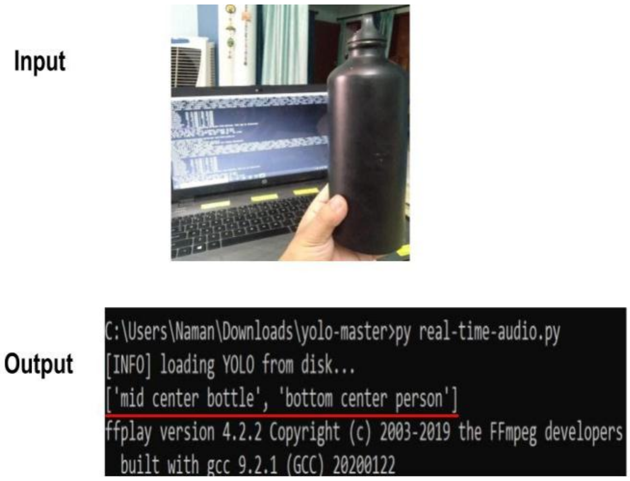


Figure 3: Hardware wiring of the proposed system

The experiment on the proposed system was conducted, we used multiple objects as shown in the figure below. The input to the system was a bottle and the system immediately detects it and converts it into audio.



```
C:\Users\Naman\Downloads\yolo-master>py real-time-audio.py
[INFO] loading YOLO from disk...
['mid center bottle', 'bottom center person']
ffplay version 4.2.2 Copyright (c) 2003-2019 the FFmpeg developers
  built with gcc 9.2.1 (GCC) 20200122
```

Figure 4: Working of the system which is detecting the bottle and the person holding the bottle.

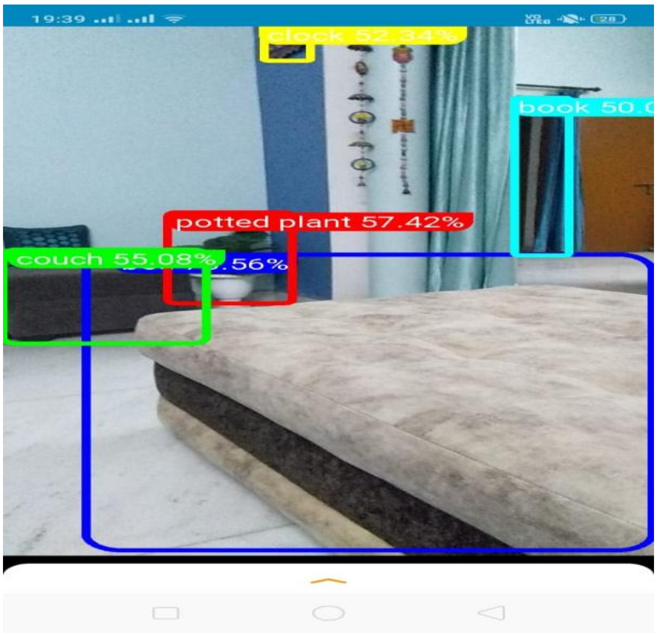
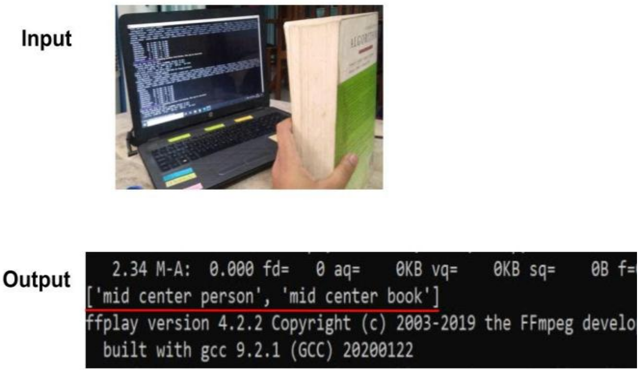


Figure 5: Objects detected by the proposed system



```
  2.34 M-A:  0.000 fd=   0 aq=    0KB vq=    0KB sq=    0B f=
['mid center person', 'mid center book']
ffplay version 4.2.2 Copyright (c) 2003-2019 the FFmpeg develo
  built with gcc 9.2.1 (GCC) 20200122
```

Figure 6: Working of the system which is detecting the book and the person holding the book.

Future Scope

This project is for the blind people who are incapable to see this colorful and beautiful world, our initiative will support them to have a better life. By this project one will be able to understand what object is present in front of him and by continuous research and development our team will be able improve this product by feeding more data to the Deep Learning algorithm by which the accuracy of the model will increase as well as the power of the algorithm to recognize more objects will increase. The object recognition system can be applied in the area of surveillance system, face recognition, fault detection, character recognition etc. The objective of this thesis is to develop an object recognition system to recognize the 2D and 3D objects in the image. The performance of the object recognition system depends on the features used and the classifier employed for recognition. This research work attempts to propose a novel feature extraction method for extracting global features and obtaining local features from the region of interest.

Conclusion

Both Deep Learning and the Raspberry technology gave us the capacity to develop these projects that will be of real benefit to the individual in need. Our system will help the visually impaired, by using research paper and based on experimental results we will be able to detect objects more accurately and independently identify objects with the exact location of an object in the x, y axis image. Which is later on converted to mp3 using gTTs and used by visually impaired person to get the exact location of the objects in the area. In our project we ha have developed a very low cost device which can be really helpful for the respective needy person

REFERENCES

[1] Kedar Potdar, Chinmay Pai and Sukrut Akolkar, ”A Convolutional Neural Network based Live Object Recognition System as Blind Aid”, arXiv:1811.10399v1 [cs.CV] 26 Nov 2018 https://arxiv.org/pdf/1811.10399.pdf

[2] Liam Betsworth, Nitendra Rajput, Saurabh Srivastava,and Matt Jones. Audvert: Using spatial audio to gain a sense of place. InHuman-Computer Interaction–INTERACT 2013, pages 455–462. Springer, 2013

[3] Evanitsky, Eugene. ”Portable blind aid device.” U.S. Patent No. 8,606,316, 10 Dec. 2013.

[4] A.Culjak, D.Abram, T. Pribanic, H. Dzapo and M. Cifrek, A brief introduction to OpenCV,"2012 Proceedings of the 35th International Convention MIPRO, Opatija, 2012, pp. 1725-1730.

[5] Rahul Kumar and Sukadev Meher, “Assistive System for Visually Impaired using Object Recognition, M.Sc. Thesis at Department of Electronics and Communication Engineering, National Institute of Technology Rourkela, Rourkela, Odisha-769 008, India, May 2015.

[6] Khushboo Khurana, Reetu Awasthi, 2013, “Techniques for Object Recognition in Images and Multi-Object Detection,” International Journal of Advanced Research in Computer Engineering & Technology.

[7] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, “You Only Look Once: Unified, Real-Time Object Detection”, University of Washington, Allen Institute for AI , Facebook AI Research, 2016.

[8] J. Redmon and A. Farha, Yolov3: An incremental improvement. arXiv, 2018.

[9] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 6517–6525. IEEE, 2017.

[10] ABradski, Gary, and Adrian Kaehler. "OpenCV." Dr. Dobb’s journal of software tools 3 (2000).

[11] Learning OpenCV 3by Gary Bradski, Adrian Kaehler Publisher: O'Reilly Media, Inc. Release Date: December 2016.

BIOGRAPHIES

Naman Mishra
Undergraduate at Babu Banarasi Das Northern India Institute of Technology, Lucknow , Uttar Pradesh, India.

Deepankar Singh
Undergraduate at Babu Banarasi Das Northern India Institute of Technology, Lucknow , Uttar Pradesh, India.