Talking Camera for Visually Impaired – A Review

[1]Shital Patil, [2]Aishwarya Raut, [3]Sagar Jaiswal

[1]Professor, [2,3]Final Year Student
Department of Computer Science,
Suman Ramesh Tulsiani Technical Campus FOE, Pune, India

***Abstract*: It becomes difficult for blind people to do their daily work independently, but an enhancement in technology can help them in some way. In this context, the present work will focus on real-time recording to speech for the blind people. The purpose of this project is to make blind people do their work independently. This system will basically identify the objects around the user and convert the data in speech finally conveying the results to the user. To achieve this, we are using text to speech and Convolutional neural network integrated with YOLO architecture, which will ultimately enable the user to take live recording and detect the objects and hear text about them.**

***Index Terms*: Artificial Intelligence, Text-to-Speech, Machine Learning, Computer Vision, Visually Impaired, Object Detection, NLP**

I. INTRODUCTION

Millions of people live in this world with incapacities of understanding the environment due to visual impairment. Although they can develop alternative approaches to deal with daily routines, they also suffer from certain navigation difficulties as well as social awkwardness. [1]. It is very difficult for blind people to do daily tasks as they are unable to perform visual tasks. Also, blind people cannot recognize the objects around them in the surrounding. Hence, using the enhancement in technology by developing a mobile application that can indeed perform the live recording to speech conversion, may it be any object, or in another support, has a great potential and utility. The technology of Convolutional neural networks enables the recognition of objects from live viewed data. This technology can be widely used for different aspects and the system developed is going to be portable and handy as well.

The technology of speech synthesis (TTS) enables a text to be played through an audio system, synthesized into human voice. The objective of the TTS is the automatic conversion of sentences, into the spoken form of the same text, in the language native to the user. This technology runs hand in hand with Natural language processing (NLP). This technology has had noticeable progress over the last few years, with many systems being able to generate a synthetic speech very close to the natural voice [3]. Due to the increasing importance of speech synthesis in many new applications, research in this area has increased considerably.

This paper describes the process of developing a mobile application, which can be used on an android phone, which allows a blind user to use the device camera and to get the reading of the existing surrounding in the captured live recording. The system uses CNN and TTS, combining them in a way that, taken together, they can provide the desired results.

II. LITERATURE SURVEY

For the development of prior work in object detection for visually impaired, we implement a system where the user can use an app which uses the camera of their android phone and predicts real-time object around them.

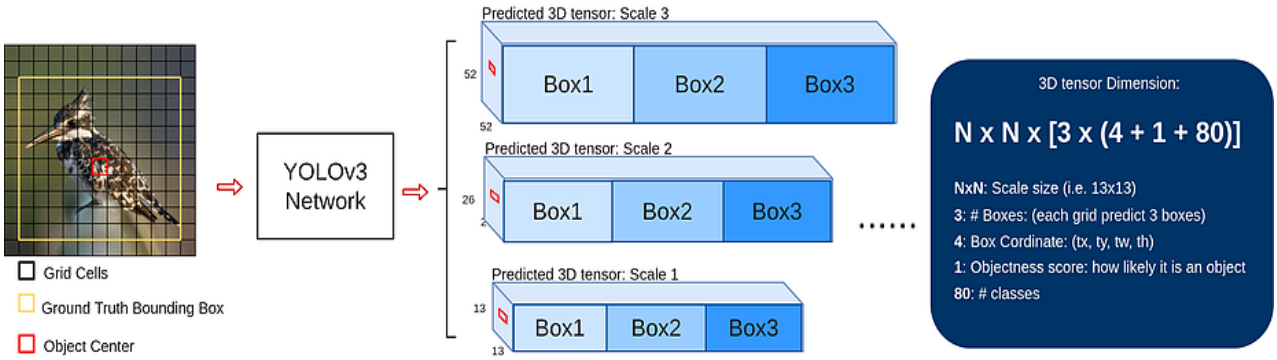
The problem of classifying these objects in real time is done by multiclass classification. Our goal is to predict these objects as fast as possible and convey the objects present to the user using TTS. To improve detection time, we are using the YOLOv3 model as it provides good fps at an average accuracy, which is needed for real-time object detection. There are other models like SSD and R-CNN which provide better mAP but the inference time are high which is not suitable for the real-time purpose. The Android application will contain the pre-trained model which will be used for detecting objects

Yolov3 Architecture

YOLOv3 is the incremental version of the basic YOLO model that has many tricks and capable of detecting small objects in an image. YOLO model has a good accuracy for real-time object detection giving an FPS greater than other models.

It is a feature-learning based network that adopts 75 convolutional layers as its most powerful tool. No fully-connected layer is used. This structure makes it possible to deal with images of any size. Also, no pooling layers are used. Instead, a convolutional layer with stride 2 is used to downsample the feature map, passing size-invariant feature forwardly. In addition, a ResNet-alike structure and FPN-alike structure is also a key to its accuracy improvement.

Figure 1. YOLOv3 process flow



Source: https://www.cyberailab.com/home/a-closer-look-at-yolov3

First, during training, YOLOv3 network is fed with input images to predict 3D tensors (which is the last feature map) corresponding to 3 scales. The three scales are designed for detecting objects of various size. Here we take the scale 13x13 as an example. For this scale, the input image is divided into 13x13 grid cells, each grid cell corresponds to a 1x1x255 voxel inside a 3D tensor. Here, 255 comes from (3x(4+1+ 80)). Values in a 3D tensor such as bounding box coordinate, objectness score, and class confidence. [5]

Second, if the center of the object's ground truth bounding box falls in a certain grid cell, this grid cell is responsible for predicting the object's bounding box. The corresponding objectness score is "1" for this grid cell and "0" for others. For each grid cell, it is assigned with 3 prior boxes of different sizes. What it learns during training is to choose the right box and calculate precise offset/coordinate. It only chooses the box that overlaps ground truth bounding box most. [5]

K-mean clustering is used to classify the total boxes from COCO dataset to 9 clusters before training. This results in 9 sizes chosen from 9 clusters, 3 for 3 scales. This prior information is helpful for the network to learn to compute box offset/coordinate precisely because intuitively, bad choice of box size makes it harder and longer for the network to learn.

Text-to-Speech

Voice synthesis, defined as TTS (an acronym for Text-To-Speech), is a computer system that should be able to read aloud any text, regardless of its origin [3]. The use of text to speech specifically aims to artificially produce human voice. It produces intelligible and natural voice result which is indeed not an easy process and contains complex algorithms wherein this concept is called Voice synthesis. TTS synthesis makes use of techniques given by NLP (Natural Language Processing). The text to be synthesized must first be processed since it is the first entry of the system. The quality of a speech synthesis can be determined by how natural it is nature.

III. Conclusion

This paper presents the development of the project Talking Camera for Blind People, considering CNN and TTS stages, to create an application that was gradually improved and refined over the project. An analysis was made regarding the CNN and TTS technologies that were used in the development of the application, in order to know the methods behind those and to understand in greater detail the mechanisms that perform object recognition and speech synthesis of texts. The project consisted of the construction of an application composed by several parts, integrating the system of live video recording by the mobile device, which is used by a CNN technology for recognition of its objects, which is then synthesized through a process of TTS. Optimizations carried out for improving outcomes resulted in a more efficient application, capable of responding to the challenge set by the theme of the project: Talking camera that speaks out what it is in surrounding for the blind. To improve the quality performance of the system, instead of using CNN throughout the image we use YOLOv3 which create multiple bounding boxes giving a better score of outcome and faster results as well. Another optimization applied to the project was using BLOB, wherein the image which is observed is converted into a grey scale which results to image enhancement using BLOB which is ANN algorithm for object detection and text to speech conversion.

The final result achieved is not perfect, since it has shortcomings regarding the recognition of objects in real time because the recognizing process of large and complex texts is sometimes slow. Also, the fact that the process of capturing the live recording images does not provide an automatic system to aid the user to orient the image capture as correctly as possible presents itself as a limitation of this application since the application is aimed at blind people, who may have more difficulties in accomplishing this task. However, this limitation can be improved in future work. The research, implementation and optimization developed allowed the design of a free application that is already in a state of possible use, even with the limitations referenced above, allowing to better understand the surroundings, provided that the light conditions are the ideal for image recording and the equipment is properly directed to object you want to listen to so that recognition and reading are as satisfactory as possible.

IV. Future Work

The robustness of the model needs to be enhanced in order to get good accuracy with image variations of brightness/contrast/blurriness, etc. Another thing that can help is to manually add image variations to input image set such that the model is less sensitive to the image variations [2]. There should be support for other languages which will be another issue to be

reviewed. At this stage, the project was designed and developed for the English language, but in the future, it should allow use in other languages, thus extending the number of people who can benefit from the advantages of the application [3]. It is also reserved for future work the increment of the menu options, allowing the user to change other application options such as the language or the setting of sounds and orientations.

REFERENCES

[1] Rui Jiang and Qian Lin, Let Blind People See: Real-Time Visual Recognition with Results Converted to 3D Audio.
[2] Zhongjie Li and Rao Zhang, Object Detection and It's Implementation on android devices
[3] Roberto Neto and Nuno Fonseca, Camera Reading for Blind People, HCIST 2014
[4] Joseph Redmon, Ali Farhadi, YOLOv3: An Incremental Improvement
[5] A Closer look at YOLOv3, Retrieved November 14, 2018, from https://www.cyberailab.com/home/a-closer-look-at-yolov3