Seer - A Computer Vision and Machine Learning Based Device for Visually Impaired

Muhammad Abdullah, Muhammad Awais Ismail, Muhammad Mehmood Ahmed, Saad Ali, Dr. Kashif Javed
Department of Electrical Engineering, University of Engineering and Technology, Lahore
abdullah612@outlook.com, awaisismail65@gmail.com, mehmoooda946@gmail.com,
saadali1906@gmail.com, kashif.javed@uet.edu.pk

***Abstract*—Vision is the most important and primitive tool for mankind to learn and interact with the environment. Sadly, there are millions of people in the world who have to live their lives in eternal darkness or with some sort of visual impairment. They have to rely on their family to fulfill their daily needs. We came up with a solution which can make the visually impaired people more independent in their daily chores by enabling them to recognize common objects in their line of sight. We want to allow them to identify familiar faces, everyday objects and recognize text that they come across in their daily life. For object detection, we are using a pre-trained MobileNet-SSD model, a Convolutional Neural Network based model, which has been trained on a huge dataset of common objects. We have fine-tuned it with our custom dataset. The custom dataset spans around 30 categories with 40 to 50 images per category. With a train/test split of 80/20, we've achieved an accuracy of around 80% for object detection. For text recognition, the text portion of the input image is extracted by a deep neural net called EAST and then an OCR named Tesseract converts it into text from image with an accuracy of around 90%. A combination of HAAR and HOG Classifier is being used to detect the faces while Nearest Means Classifier employing the vector embeddings created from our own custom dataset is being used to recognize them. The collected data contains around 10 persons with 50 images per person. With a train/test split of 80/20, we've achieved an accuracy of around 85% for face recognition. The major tools being deployed are Python, Colab, Numpy, Tensorflow, OpenCV and ImUtils.**

***Index Terms*—Computer Vision, Convolutional Neural Net, Dataset, Detection, Recognition.**

I. Introduction

On Earth, 285 million people are visually impaired, out of which 39 million people are living in complete darkness i.e are totally blind and 246 million people are suffering from some sort of vision deficiency which may be mild or severe. These numbers will keep on increasing, up-to nearly 75 million blind and 200 million people with visual impairment by the year 2020. As it is hard to identify different things for visually impaired people like objects, text and faces, our device will help them by using speech to identify these products.

Visual aids for blind people are being researched by a lot of computer vision experts. Researchers are using feature extraction and matching to detect and recognize objects in images. They are using the concept of local feature extraction [1]. New models for object detection are being developed which are getting faster and more accurate with passage of time. Faster and faster feature extractors like MobileNet are appearing which combine with efficient classifiers like SSD to produce excellent results [2]. A new method which includes combining of complex classifiers in cascade is also developed which allows background regions of the image to be quickly discarded, called the HAAR Cascade [3]. Using locally normalized histogram of oriented gradients (HOG) features similar to SIFT descriptors in a dense overlapping grid gives very good results for person detection [4]. EAST detector is a pipeline that directly predicts words or text lines of arbitrary orientations and quadrilateral shapes in full images, eliminating unnecessary intermediate steps, with a single neural network [5].

II. Problem Description

Detection and recognition of nearby objects is a big problem for visually impaired. They have to either touch the object or get some help to tackle these problems. Visually challenged people also find it impossible to read text or recognize someone's face.

There are many projects which are targeted at navigation for the blind. Image of surrounding is converted into auditory stimulation or a vibratory mechanism. These projects are tackling the issues of detection but they fail to recognize/identify the objects. To a computer, one object can appear in multiple forms due to change in angle, difference in illumination, different cameras etc. These problems increase the difficulty of object identification by a lot.

Therefore, a system must be designed to overcome these issues to allow visually impaired people to recognize objects, faces and text at different angles and conditions.

III. Methodology

We are taking input image from a camera and feeding it into the neural net after processing it. The neural net performs calculations and output layer neurons are activated with different confidence.

A. Image Input

First of all, the image is captured with the help of a camera. The image is then sent to the system for further processing and detection.

B. Image Processing

After an image is captured, it must be re-sized to meet the requirements of our net.Various algorithms and methods are used for the purpose of Image processing and detection

[6], [7]. The image pixel matrix is in RGB format which is converted to BGR. Similarly, grey scaling, mean subtraction is performed before the image passes through the neural net.

C. Machine Learning

We are using Convolutional Neural Nets to detect and recognize objects, faces and text after processing the image.CNNs are primarily used to solve difficult image-driven pattern recognition tasks and with their precise yet simple architecture, offers a simplified method of getting started with ANNs [8].

1) Object Recognition: MobileNet is being used as feature extractor for object detection. We are using SSD on top of it to detect/recognize objects. MobileNet-SSD v2 provides a good trade-off between speed and accuracy. It was trained on our custom dataset spanning across 30 objects with about 50 images per object.

2) Facial Recognition: In this part, faces of around 10 different persons are first detected and then recognized. The model being used is a combination of HOG and HAAR algorithms. It was trained on our custom dataset spanning across 10 persons with about 50 images per person.

3) Text Recognition: In this part, text appearing in English language is first detected and then recognized. The model being used is a combination of EAST and Tesseract OCR.

D. Speech Synthesis

There are different approaches for speech synthesis or TTS (Text to speech) conversion. Both open-source and proprietary softwares for TTS are available, but we opted for the open-source platform. By bundling TTS engine with the image processing module, a complete system is made [9].

E. Audio Output

After object detection through image processing and text to speech synthesis, the final output is conveyed to the user in the form of audio.

IV. Implementation

Three modes are being implemented in the device. These modes switch between models for text recognition, facial recognition and text recognition.

A. Software

We are using Python language which is being widely used for machine learning. OpenCV is being used to created a BLOB (Binary Large Object) from image. BLOB contains large binary objects while the small objects are filtered out as noise. OpenCV is also being used to load our models into memory. The BLOB is an input to our models which detect objects and classify them.

ImUtils is being used to capture frames from camera. The captured frame passes through the respective model depending on the user choice after BLOB extraction. The user choice of model is detected by interrupts through buttons for each respective model. The output of the model includes classifications alongwith respective confidence. The classifications with low confidence are filtered out while high confidence classifications becomes an argument to Python TTS library to produce an audio output.

B. Hardware

The complete hardware is divided into two distinct portions.

The first portion of hardware (Figure 1) consists of a Raspberry Pi 3B+ microcontroller enclosed in an acrylic case. The casing is further embedded with a cooling fan on the top and heat sinks on controller chips for the purpose of cooling. Furthermore, a single power switch for the purpose of turning the hardware on and off is installed. Three push buttons are also installed with the help of jumpers for the purpose of interrupt generation in the context of mode switching. An indicator LED for power is also installed. Connection between power bank and microcontroller is established through a USB cable and a switch which work as a power source.

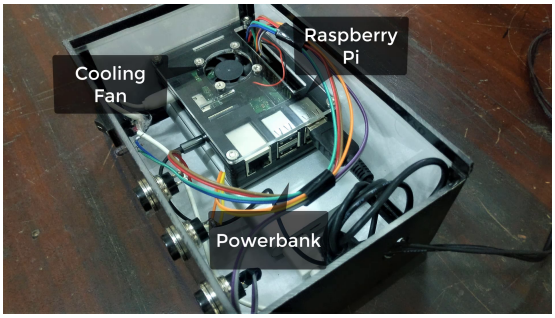

Fig. 1. Hardware Portion 1

In the second portion of hardware (Figure 2), a small webcam and a set of earphones are fitted on a standard pair of glasses, and further connected to microcontroller in the first hardware portion via cables.

Fig. 2. Hardware Portion 2

The final hardware consists of an acrylic box containing the microcontroller and power bank along with three push buttons and a power switch, and a pair of glasses with webcam and earphones attached to it.

V. Results and Discussions

Hardware testing was done for object recognition, facial recognition and test recognition for their respective accuracy and speed. We also tested our power bank to test the amount of time to drain the battery.

A. Object Detection

The model deployed on hardware for object detection and recognition was MobileNet-SSD v2. The accuracy achieved was around 80% while the speed reached around 1.8 fps. A sample image is shown in Figure 3.

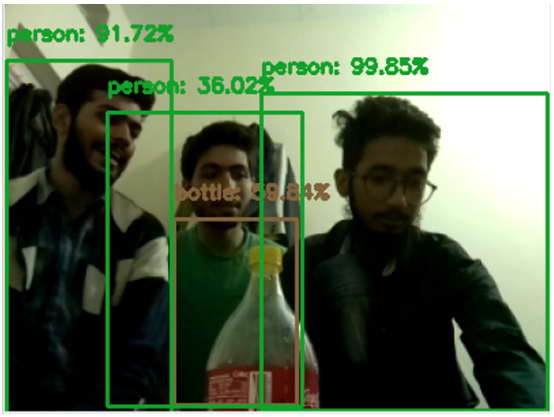

Fig. 3. Hardware Results for Object Recognition

B. Facial Recognition

The model deployed on hardware for facial detection and recognition was HOG and HAAR combination. The accuracy achieved was around 85% while the speed reached around 1.6 fps. A sample image is shown in Figure 4.

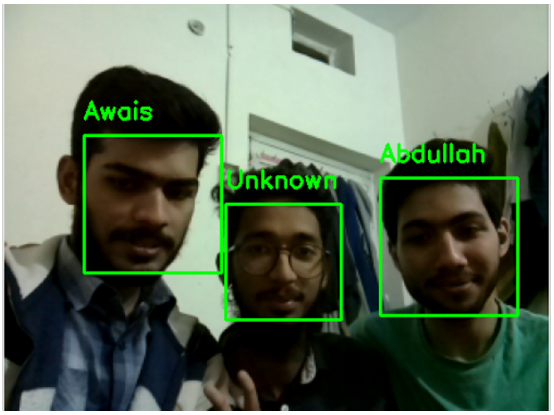

Fig. 4. Hardware Results for Facial Recognition

C. Text Detection

The model deployed on hardware for text detection and recognition was EAST and Tesseract combination. The accuracy achieved was around 90% while the speed reached around 1.4 fps for just detection while it took around 4 to 5 seconds to complete recognize the text in the image and this time depends upon the amount of text in the image. A sample image is shown in Figure 5.

Fig. 5. Hardware Results for Text Recognition

D. Power Usage

The battery time was tested with 10000 mAh power bank. The current usage of Raspberry Pi is 0.5 amps in idle condition which peaks at 1.2 amps at max usage. This resulted in weighted average of around 0.7 amps. These values gave a total battery time of around 10 to 12 hours.

VI. Conclusion and Future Work

A. Summary

The purpose of the project is to help the people with visual disabilities to recognize everyday objects, faces and textual information in their line of sight to make them more independent in their daily chores. The accuracy achieved for object recognition is around 80%, for facial recognition is around 85%, and for text recognition is around 90% as shown in Figure 6. The solution is cheap, compact and wearable being fairly quick and accurate. However the speed and accuracy could be improved even further.

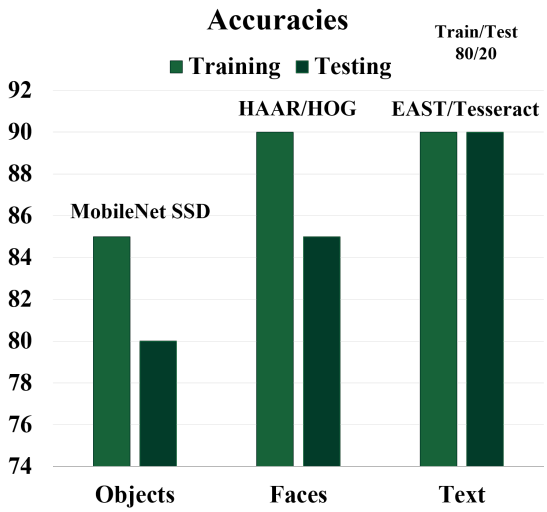

Fig. 6. Graph for Accuracies

TABLE I
COST ANALYSIS

Item	Cost (PKR)
Raspberry Pi	6000
Powerbank	2000
Webcam	3000
Earphones	400
Glasses	200
Cases/Boxes	1000
Misc	400
Total	13000

B. Applications

The technology can be useful in many other applications aside from helping the blind.

- The technology can be used in self-driving cars to detect other cars, identify pedestrians and read road signs.
- The technology can also be useful in surveillance procedures to monitor individuals and their activities.

C. Future Prospects

Following are some of the recommendations for future work:

- The number of objects or known faces can be increased by using more data.
- The ability to recognize languages other than English can be added as well.
- The accuracies can be increased by using a larger dataset comprising of more data.
- The speed can be increased by using a more powerful microcontroller or computer.

D. Cost Analysis

A comprehensive final cost analysis of the project and the product is given in Table I.

REFERENCES

[1] H. Jabnoun, F. Benzarti and H. Amiri, "Object Detection and Identification for Blind People in Video Scene", *15th International Conference on Intelligent Systems Design and Applications (ISDA)*, 2015.

[2] W. Liu, D. Anguelov, et al., "SSD: Single Shot MultiBox Detector", *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[3] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", *Conference on Computer Vision and Pattern Recognition*, 2001.

[4] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection", In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005.

[5] X. Zhou, C. Yao, et al., "EAST: An Efficient and Accurate Scene Text Detector", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[6] K.M.M. Rao, "Overview of Image Processing", In: Readings in Image Processing Fundamentals Of Digital Image Processing, Prentice-Hall, 1989.

[7] E.L. Hal, "Computer Image Processing and Recognition", Academic Press, 1979.

[8] K. OShea and R. Nash, " An Introduction to Convolution Neural Networks", 2015.

[9] Hariprasad, K.C., et al., "Information awareness for the visually-impaired using machine-vision", In: Advances in Natural and Applied Sciences, 2017, 11(6), 220-225.