THIRD EYE: AI BASED VISION SYSTEM FOR VISUALLY IMPAIRED USING DEEP LEARNING

Abstract

The objective of the project is to design and build a vision-based AI system that leverages deep learning techniques for helping visually impaired and blind persons. Individuals who have lost their vision, as well as their families, friends, and society, are all impacted. Complete vision loss or degradation can be frightening and overwhelming, causing those affected to doubt their ability to maintain their independence, pay for necessary medical treatment, keep their jobs, and provide for themselves and their families. Loss of vision has far-reaching health implications that extend beyond the eye and visual system. Falls, injury, and deterioration in mental health, cognition, social function, employment, and education levels have all been linked to vision loss. The project aims at providing vision-based solution for visually impaired using state-of-the-art deep learning techniques.

Keywords: Machine Learning, Deep Learning, Computer Vision, Face Identification, Google Cloud Vision, Classification, Object Detection.

Authors

Ancy Thomas
Assistant Professor
Department of Computer Science and Engineering
Acharya Institute of Technology
Bangalore, India

Shyam U
Student
Department of Computer Science and Engineering
Acharya Institute of Technology
Bangalore, India

Shreyon Barman
Student
Department of Computer Science and Engineering,
Acharya Institute of Technology
Bangalore, India

I. INTRODUCTION

1. **Description of the project:** This project aims to aid blind people by identification of text on physical surfaces and recognizing faces. In this project, artificial intelligence is used to provide vision to the visually impaired. In order to aid the blind in simple daily actions, we are attempting to build an Artificially Intelligent system. Essentially, the system consists of four modes: Recognizing mode, Reading mode, Detection mode, and Currency Identifier mode.

    - **Recognizing mode:** Through speech input to the AI system, the user directs the system to recognize or train an individual. Here the system captures the image of the person and trains the model to recognize the individual when encountered in future.

    - **Detection mode:** Of the three modes, this will be the most active. When the person is within the frame range of the system, the individual that was trained/recognized in the Recognizing mode will be detected by the detection model and voice output will be given by it.

    - **Reading mode:** A primary objective of the project is to provide assistance with reading to the blind. When a user activates Reading Mode, the system will capture the image of the page or text that is presented to it and read the entire text for that user. This will be especially helpful to visually impaired individuals.

    - **Currency identifier mode:** When the user activates the Currency Identifier mode by speaking and places a currency note in front of the frame, the system captures the image of the currency note and outputs voice output stating the currency value.

2. Objective of the project

This project aims to carry out the following objectives:

- Assisting the visually impaired people in making their back-breaking lives comfortable.
- Recognition of the printed text and read it aloud
- Recognize a particular person coming in the camera frame and spell out their name.
- Identification of Indian Currency notes for day-to-day transactions

II. PROBLEM STATEMENT

1. **Problem statement:** To design and build a vision-based AI system that leverages deep learning techniques for visually impaired and blind persons

2. Related works

- In [1], the authors have designed and developed a system which uses an OCR reading system that converts text available in an image to speech/voice output. The system provides a helping hand to blind people in reading the text. But the implementation specified in the paper is through a mobile application designed for the Android platform. This makes its usage difficult for the blind people to navigate the app.

- In [2], the authors have proposed a system that capture images when pointed by the user and locates any text in the image. The identified text is then converted into an audio output. It uses a single and light weight neural network thereby achieving great efficiency in both performance and speed. However, the system can make the blind person's job of positioning and orienting the camera difficult as tracking the camera over the lines of text is needed.

- In [3], the authors have proposed an approach for text localization and extraction in order to detect text areas in images using OCR and CRNN. The system has an advantage of improving the accuracy of traditional OCR using CRNN. However, capturing all the text in an object could be difficult to a blind person through the application.

- In [4], the authors have implemented a facial and hand gesture recognition system. Haar Cascade classifiers has been used for facial detection and LBR histogram has been used for identification of a person. Face identification and detection is processed in real time which helps the blind person recognize the person coming in the range of camera. However, additional features can be included in addition of face recognizer.

- In [5], the authors have developed a guidance system that captures images using a smart glass paired with sensors. The system's processor extracts the detected objects from the images and provides speech-based output. The developed system is capable of efficient detection of objects in the camera frame using YOLO algorithm and Google Vision API. But the developed system could be expensive due to the integration of a processing unit and a camera into a smart glasses.

3. **Significance of the project:** Individuals who suffer from vision loss, as well as their families, friends, and society, are all impacted. Complete vision loss or degradation can be terrifying and overwhelming, leaving those affected to question their capacity to preserve their independence, pay for necessary medical treatment, keep their jobs, and provide for themselves and their family. Vision loss has far-reaching health repercussions that go beyond the eye and visual system. Vision loss has been linked to falls, injury, and worsened status in domains spanning mental health, cognition, social function, employment, and educational achievement. [6] Vision loss has a significant economic impact. Direct medical expenses, other direct expenses, lost productivity, and other indirect costs for visual disorders across all age groups totaled $139 billion in 2013 dollars, according to a national study commissioned by Prevent Blindness [7], with direct costs for the under-40 population reaching $14.5 billion dollars [8]. These costs have an impact on not only national health-care spending but also associated expenses and individual and family resources. Our project aims to mitigate the impact of vision loss by providing features such as recognition of text, facial detection and identification and currency note detection.

III. PROPOSED METHODOLGY

1. **Proposed architecture:** The problem statement has been decomposed into 4 modules – Currency Identifier Mode, Recognizing mode, Detection mode and Reading mode. The system receives a speech input to determine which mode will handle the user's request.

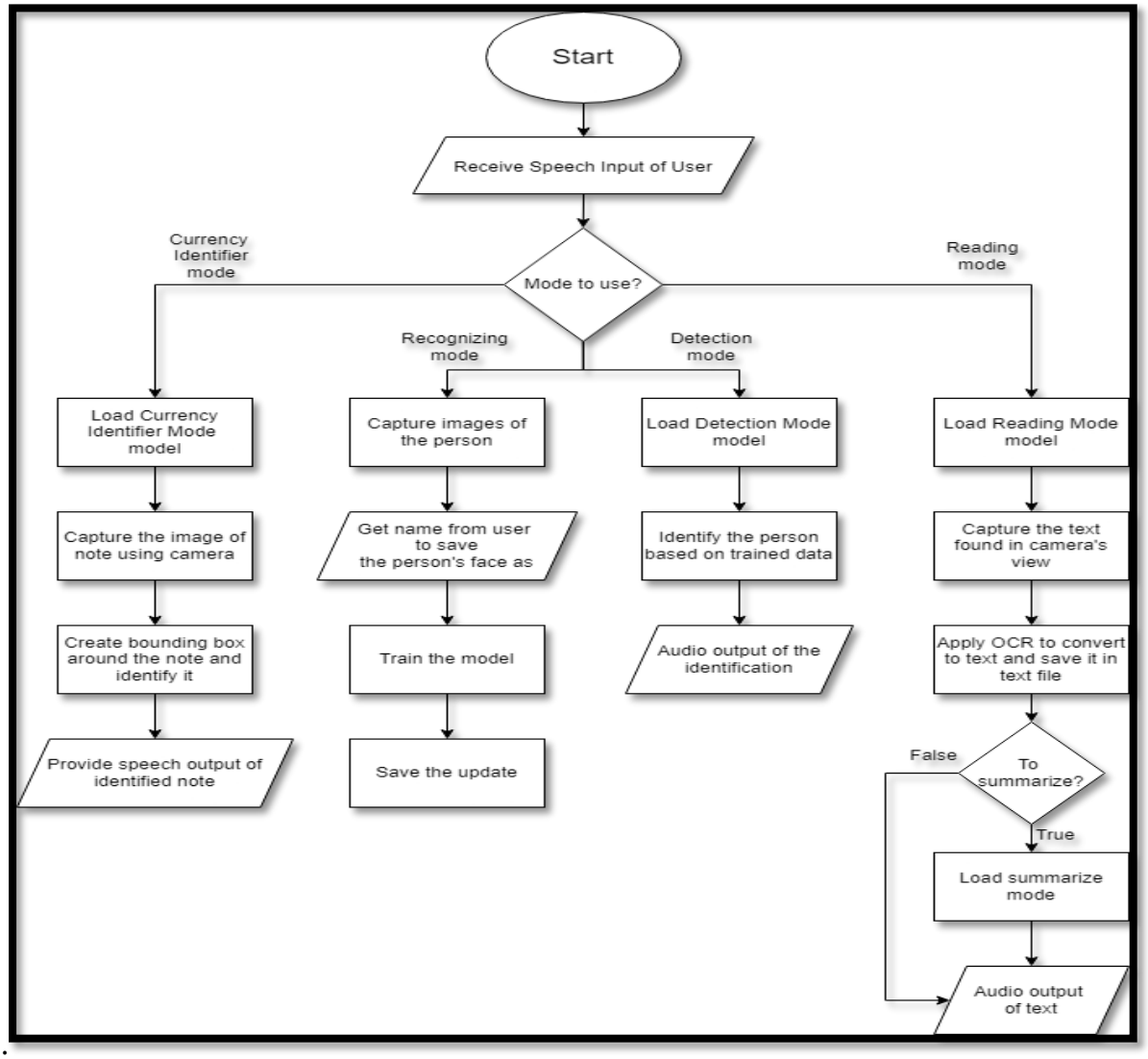

Figure 1: Architecture Followed in Methodology

2. **Proposed modules:** The modules identified from the problem statement are described in this section. Each module is dedicated to handling a specific task and represents a mode as identified by the problem statement. As a result, the system can use all of the modules listed below to assist the visually impaired user based on their speech input.

    - **Reading module:** This module is responsible for converting text from an image into a vocal output. This module detects and recognizes text on printed materials before converting it to voice output. The conversion of text in images into digital text can be done using Tesseract, an opensource OCR engine. A text-to-speech conversion program such as pyttsx3 can then utilize the digital text. The application of this module requires that the text identification be done with high accuracy. Accurate text identification ensures that voice output correctly conveys the available text to the user.

- **Currency identifier module:** This module is in charge of detecting and identifying Indian currency notes in real time when they are placed in the camera frame. It should be capable of identifying Indian currency notes worth 100, 200, and 500 rupees. The model is trained on a custom dataset that includes more than 2,000 images which were captured using a script. The dataset is then annotated using an opensource labelling software. The possible application of this module requires faster processing and detection of currency notes with accurate results. YOLOv5, the most recent version of YOLO, has been selected for this model because it outperforms most of the state-of-the-art detection algorithms present with 140 FPS.

- **Face identification module:** This module is expected to identify known people. It does so by capturing images of individuals on command and using them to train a neural network that can recognize the individual. Its goal can thus be divided into two parts: (1) it should be capable of detecting faces that appear within the camera frame whose output is used; and (2) it should be able to recognize the detected faces. The face detection module serves to capture the necessary facial details and store them. The face recognition module utilizes the stored facial details to train a neural network. It results in the neural network being capable of identifying the same person when detected at a later point of time.

- **Summarization module:** This module accomplishes the task of summarizing for the user a paragraph of text that visually impaired people could encounter on a daily basis. With the help of this function, paragraphs will be condensed into summaries with important ideas and takeaways for improved information capture via audio output. If the user requests a summary of the text that was read by the system, the summarization module is activated.

IV. IMPLEMENTATION

The proposed system is implemented by combining numerous technologies as discussed below:

1. **Reading module:** Read module is responsible for converting text from an image into a vocal output. This module recognizes text on printed materials before converting it to voice output.

   This module achieves the functionality of recognizing the text on a printed document when it is placed in front of a camera and reading it out loud to the user. This feature will make it easier for persons who are blind to find information in printed texts that will benefit them in their daily lives. Printed materials give us important information that keeps us informed and up to date. Thus, adding this feature will enhance visually challenged people's capacity to engage with their surroundings.

   This module was implemented using OCR techniques with the assistance of Google Cloud Vision API. The Google Cloud Vision API enables developers to create vision-based machine learning applications based on object detection OCR, etc. [9]

2. **Currency identifier module:** Currency identifier module is in charge of detecting and identifying Indian currency notes in real time when they are placed in the camera frame. It is capable of identifying Indian currency notes worth 100, 200, and 500 rupees.

   This module achieves the functionality of identifying Indian Currency Notes when placed in front of camera. By providing this capability, physical currency notes will be more accessible to persons who are blind or visually impaired. This will be useful for those who might not have access to or be able to use digital payment. The module needs to process and detect money notes quickly and with accurate results.

   YOLOv5, the most recent version of YOLO, has been selected for this module's implementation because it outperforms most of the state-of-the-art detection algorithms. YOLO, an acronym for 'You only look once', is an object detection algorithm that divides images into a grid system. Each cell in the grid is responsible for detecting objects within itself. It is lightweight and easy to use. Custom dataset was used and annotations was done using LabelImg software. Both annotation files and images were sent as inputs for training. Over 4,000 images and annotation files were used for training the model. [10]

3. **Face identification module:** Face identification module is expected to identify known people when seen on camera frame. For every face that has been trained embeddings of the face with their names will be saved. During detection, the face that is in the frame is compared with the embeddings and prediction is given as speech output.

   As long as their face details have been saved, this module can add a new person at the user's request and afterwards recognize that person when they appear in the camera. Visually challenged people can use this function to add people they know to the system. This feature is also capable of recognizing and identifying those have already been registered.

   This module was implemented using Multi-Task Cascaded Convolutional Neural Network (MTCNN). The framework adopts a cascaded structure with three stages of carefully designed deep convolutional networks that predict face and landmark location in a coarse-to-fine manner. The CNN consists of three stages. In the first stage, it produces candidate windows quickly through a fast Proposal Network (P-Net). Then, it refines the windows to reject a large number of non-faces windows through a Refinement Network (R-Net). In order to further refine the output and output the positions of facial landmarks, it uses the Output Network (O-Net). Thanks to this multi-task learning framework, the performance of the algorithm can be notably improved. [11]

4. **Summarization module:** The summary module's goal is to condense the text input into a concise abstract. This module handles the text identified by the read module before it is read to the user.

   The user may request to receive a quick summary of the recognized text. Distil BERT is used to implement this module. It is a pre-trained smaller general-purpose language representation model. It can be fine-tuned with good results on a variety of activities. Smaller language models pre-trained with knowledge distillation can achieve performance comparable to big size models. Knowledge distillation is a compression

method where a small model (the student) is educated to mimic the actions of a larger model (the teacher) or group of models [12].

Figure 2: Start Window

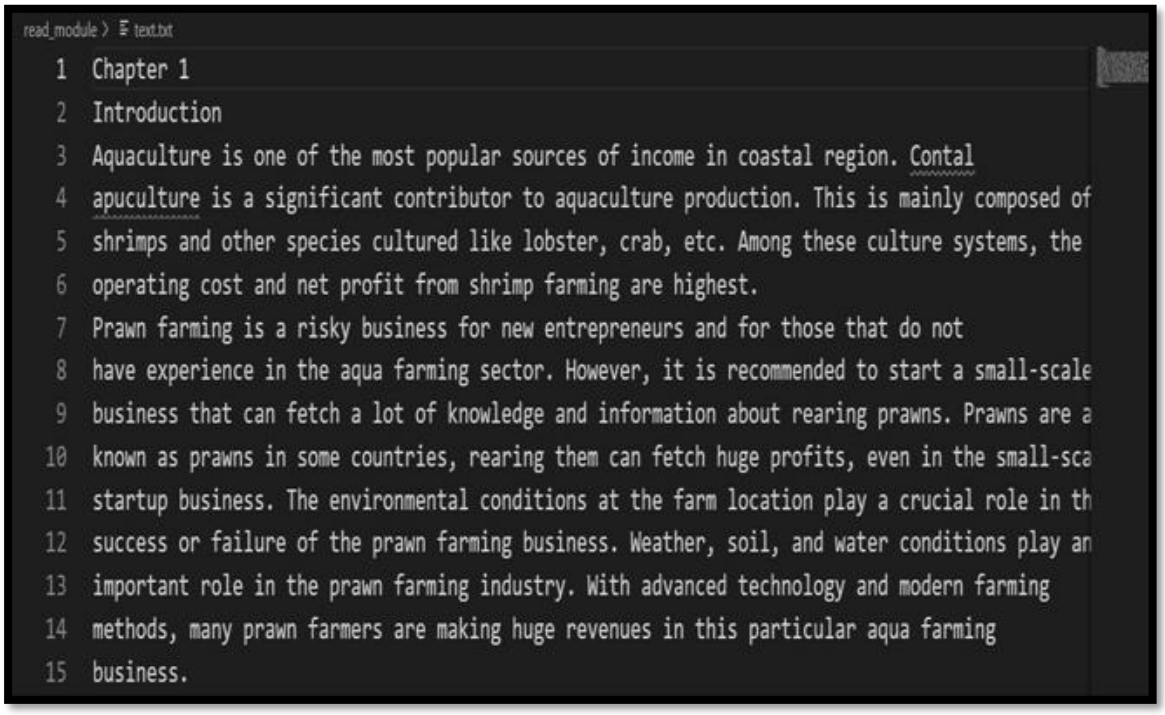

Figure 3: Detected Text from Reading Module

V. RESULTS

This project aims to create deep learning models which can assist visually impaired and blind people. The developed models are individually capable of reading text, identify trained people and identify Indian currency notes. The models are merged into a software system which is operated using voice commands.

The entry point of the system is shown in figure 2. It contains short description of all the modules with a power button. Once the power button is clicked the system gets activated enabling the users to give speech input based onthe users' requirements.

Figure 3 shows the text that was identified and extracted into a file. The same text file is then sent to speech output module and the text is read aloud. Output of the currency identification module is shown in figure 4. The images are displayed in batches as shown.

Figure 4: Currency Identification by the Model

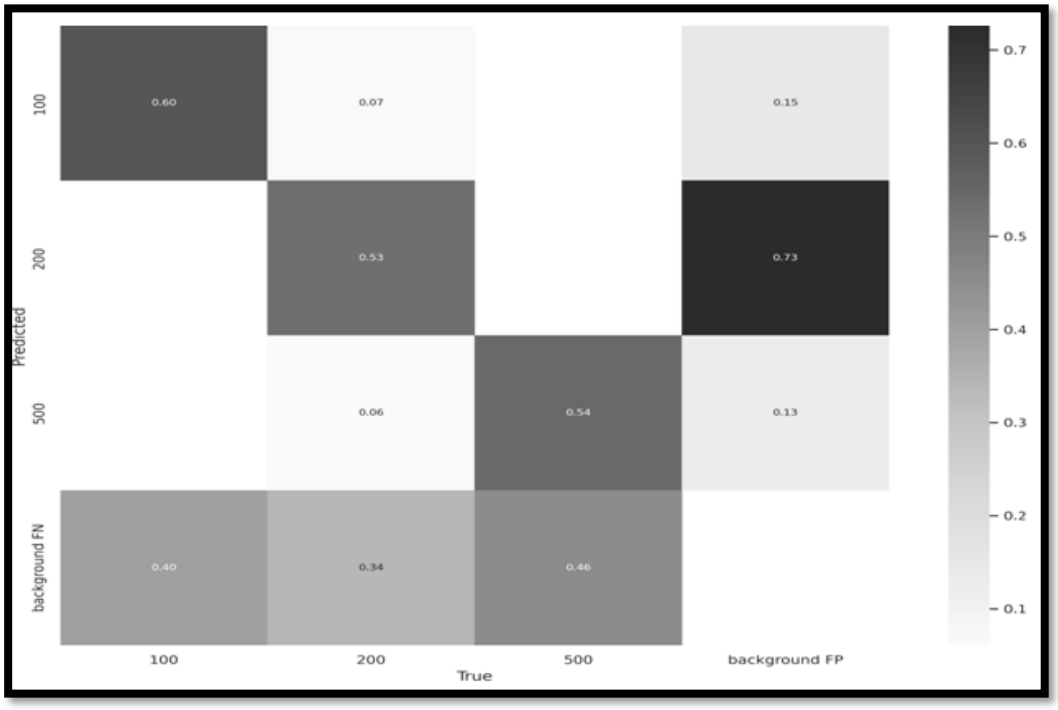

Figure 5: Normalized Confusion Matrix for Currency Identification Output

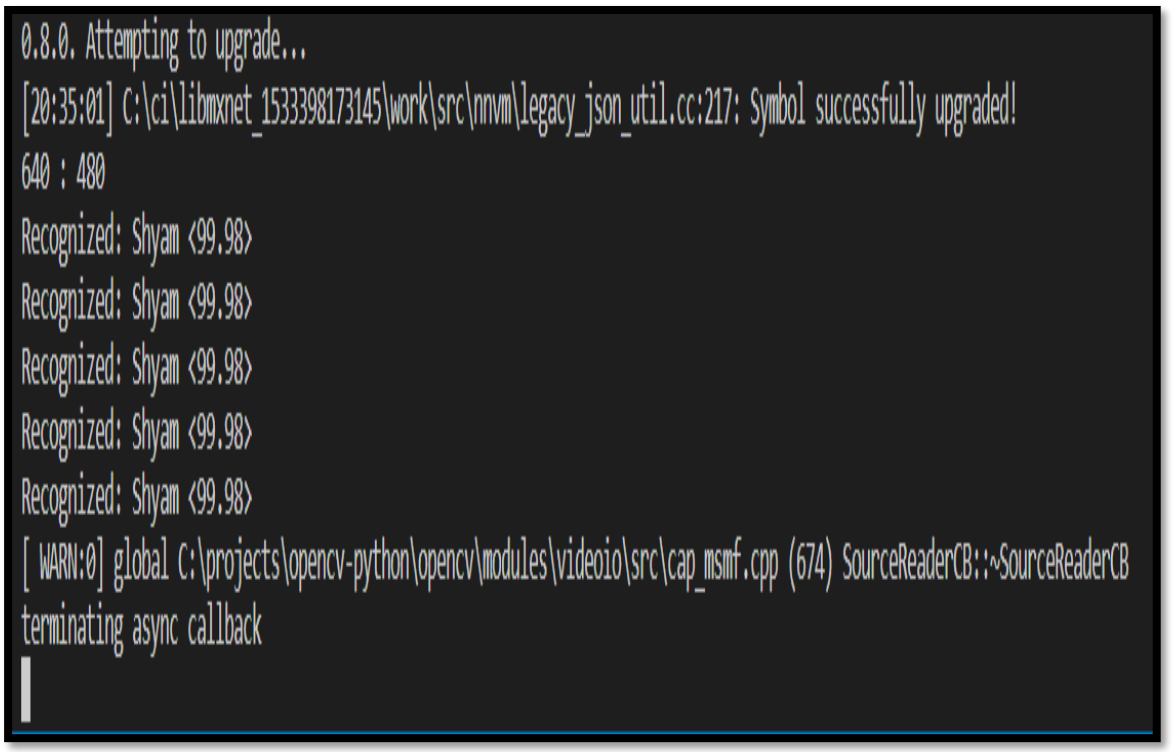

Figure 7: Image Capture for Training

The confusion matrix of the currency identification model is shown in figure 5. Figure 6 shows the output of the system when the system identifies a person in camera. Figure 7 shows capturing the images for training where only the face part is identifiedand captured for training. The training process when a new face is to be saved is shown in figure 8. After the facial landmarks are identified, a bounding box is applied whichis used to crop the image as shown in figure 9 in order to store relevant facial details and store it to a folder during the training process.

Figure 7 shows the output of the system when the system identifies a person in camera. Figure 8 shows capturing the images for training where only the face part is identified and captured for training. The training process when a new face is to be saved is shown in figure 9. After the facial landmarks are identified, a bounding box is applied whichis used to crop the image as shown in figure 10 in order to store relevant facial details and store it to a folder during the training process.

VI. CONCLUSION AND FUTURE SCOPE

This project was capable of creating models for assisting blind people. Using our system, blind people will be capable of reading texts, identify people whom they know and identify Indian currency notes for daily transactions

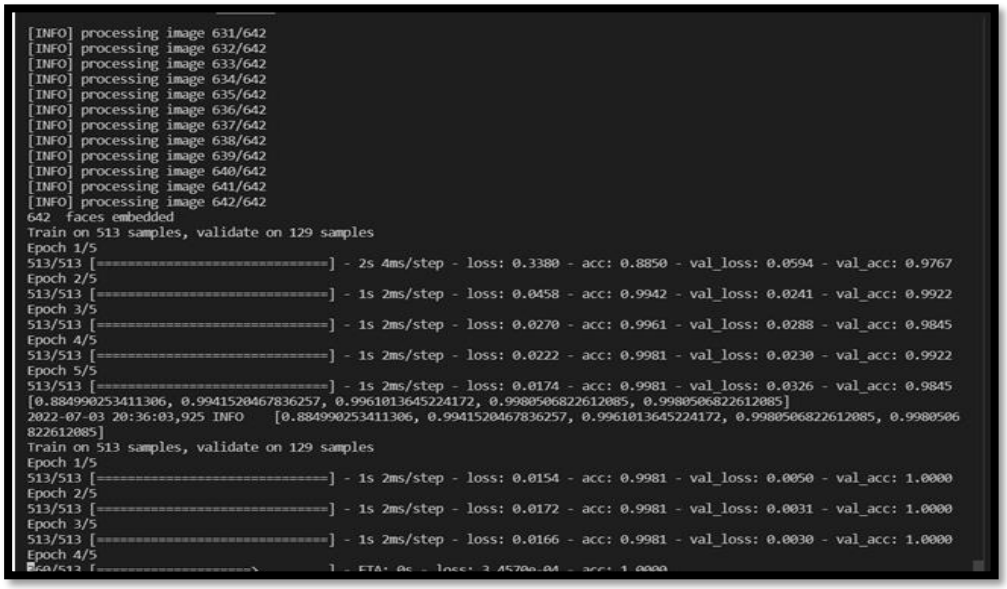

Figure 8: Training Process When a New Face is to be saved

Text on printed materials can be captured using system's camera which can then be summarized on user's request. Indian currency notes are identified when placed on the camera frame. The system is capable of identifying known person when in camera frame. It is also capable of training itself under a minute to identify new person. The system is completely handled using voice commands and outputs audio for easier usage as we are assuming the user is blind but not deaf.

The system can be put into a wearable device, such as eyeglasses, which will use the trained models to help blind people in real life scenarios. This would result in greater usability by those who are visually impaired and blind. With the press of a button on the eyeglasses and voice commands, the system would be convenient to use.

Figure 9: Cropped Image after bounding box is applied

This project can be expanded by providing more features to greatly enhance blind people's capabilities such as expanding the identification of Indian Currency notes to identification of everyday items

REFERENCES

[1] A. Mathur, A. Pathare, P. Sharma and S. Oak, "AI based Reading System for Blind using OCR," 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), 2019, pp. 39-42, doi: 10.1109/ICECA.2019.8822226.

[2] S. Kowshik, V. R. Gautam and K. Suganthi, "Assistance For Visually Impaired Using Finger-Tip Text Reader Using Machine Learning," 2019 11th International Conference on Advanced Computing (ICoAC),2019, pp. 7-12, doi: 10.1109/ICoAC48765.2019.246808.

[3] T. Shah and S. Parshionikar, "Efficient Portable Camera Based Text to Speech Converter for Blind Person," 2019 International Conference on Intelligent Sustainable Systems (ICISS), 2019, pp. 353-358, doi: 10.1109/ISS1.2019.8907995.

[4] S. Sharma, S. Jain and Khushboo, "A Static Hand Gesture and Face Recognition System for Blind People," 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN), 2019, pp. 534-539, doi: 10.1109/SPIN.2019.8711706.

[5] P. S. Rajendran, P. Krishnan and D. J. Aravindhar, "Design and Implementation of Voice Assisted Smart Glasses for Visually Impaired People Using Google Vision API," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1221-1224, doi: 10.1109/ICECA49313.2020.9297553.

[6] Welp A, Woodbury RB, McCoy MA, et al., editors. 2016 Sep 15. 3, The Impact of Vision Loss. Available from: https://www.ncbi.nlm.nih. gov/books/NBK402367/.

[7] Wittenborn J, Rein D. Cost of vision problems: The economic burden of vision loss and eye disorders in the United States. Chicago, IL: NORC at the University of Chicago; 2013.

[8] Wittenborn JS, Zhang X, Feagan CW, Crouse WL, Shrestha S, Kemper AR, Hoerger TJ, Saaddine JB. The economic burden of vision loss and eye disorders among the United States population younger than 40 years. Ophthalmology. 2013;120(9):1728–1735

[9] "Cloud Vision documentation," Google. Available: https://cloud. google.com/vision/docs.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection. arXiv, 2015. doi: 10.48550/ARXIV.1506.02640.

[11] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/lsp.2016.2603342.

[12] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, Distil BERT, a distilled version of BERT: smaller, faster, cheaper and lighter. Ar Xiv, 2019. doi: 10.48550/ARXIV.1910.01108.

