An Assistance System for Visually Challenged People Based on Computer Vision and IOT

1st Akash Bhuiyan
Arollo Tech Limited
Dhaka, Bangladesh
akash.bhuiyan@gmail.com

2nd Md Ariful Islam
Arollo Tech Limited
Dhaka, Bangladesh
arifultusharr@gmail.com

3rd Md Hasan Shahriar
Arollo Tech Limited
Dhaka, Bangladesh
hshahriar75@gmail.com

4th Tahamir Hasan Supto
Arollo Tech Limited
Dhaka, Bangladesh
tahamirhasan@iut-dkaha.edu

5th Mohammad Abul Kasem
Arollo Tech Limited
Dhaka, Bangladesh
iamrupok@gmail.com

6th Mohammad Eusuf Daud
Alo Ltd
Dhaka, Bangladesh
medaud@gmail.com

***Abstract*—This work presents a working wearable glass that can simplify some of the basic difficulties of a visually impaired person with the help of computer vision and internet of things. The glass can see the surroundings, hear voice commands, process information and send feedback to the wearer through bone conduction technology without blocking ear hole so that the wearer can be connected with the glass and surrounding world simultaneously. The glass recognizes many common objects and known human faces accurately in real time, which provides the wearer a certain degree of freedom to move alone with less fear in limited environment. The use of scientific inventions and technological advancements in developing such systems will enhance social empathy towards the least observed ones.**

Keywords—Computer Vision, Computational Intelligence, IOT, Artificial Intelligence, Image and Video Processing, Object Recognition, Facial Recognition, Bone Conduction.

I. INTRODUCTION

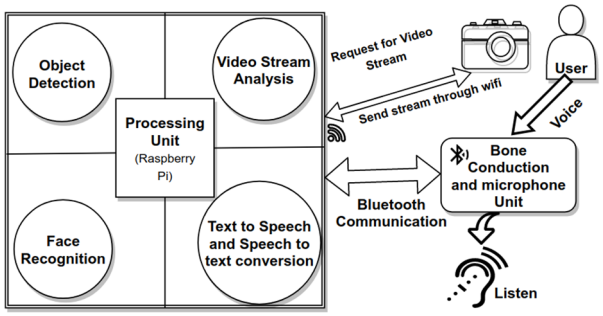
The eyes are one of the major human perceptual systems and a large share of human faces vision difficulties ranging from partial vision impairment to complete blindness. According to World Health Organization fact sheet, at least 2.2 billion people have vision impairment or blindness globally. The majority of people with vision impairment are over the age of 5 years [1]. To address this issue with a probable technical solution; 3D product design, customized hardware and processing unit, object recognition, facial recognition, speech to text and text to speech conversion, Bluetooth and Wi-Fi signal usage, bone conduction technology have been incorporated into one solution and presented as a wearable glass. Though the total working processes of the glass is complex and spread across many domains, this work is concluded by demonstrating core design and features only.

II. RELATED WORK ON VISION AID INSTRUMENT

Use of vision aid instruments exists for centuries. Reading stone was the first vision aid instrument invented around 1000 AD [3]. It was a glass sphere, laid on top of the reading object to magnify letters. The first wearable eye glasses were invented around the 13th century in Italy [3]. From those initial attempts to modern eye wears, are being used to help partial visual impaired people. But the segment for aiding completely blind or accidental blind people has been less practiced. With the advancement of medical and computational capabilities, this segment has started to see the light of hope. Many individuals or groups have demonstrated their works on smart eyewear to incorporate machine vision into human understanding to guide blind people by recognizing public signs [2], navigating to avoid obstacles. Smaller but powerful microprocessors, integrated component units and reliable power sources make on device processing possible and inspires many practitioners to work in this less cared domain.

III. ARCHITECTURAL OVERVIEW OF THE SYSTEM

The high level overview of the core system is represented in "Fig. 1(a)". It is presumed that the user of the system is a blind or partially blind person who has the ability to speak, hear and understand sound. When the user speaks to the glass, built in microphone in bone conduction unit receives the voice input signal and sends it to processing unit via bluetooth. Processing unit accepts audio signal and converts it into text format using speech to text conversion technology. Initially the system can detect and recognize English and Bengali languages only. More language support will be added in future upgrade.



(a) High level Overview of the system Architecture

(b) Front and Rear View of 3D designed eyeglass

Fig. 1. System architecture overview and two side views of the 3D Designed eyeglass.

978-1-7281-7366-5/20/$31.00 ©2020 IEEE

Based on user voice input, processing unit performs several actions. If the input is to see surrounding and recognize objects or faces, processing unit will request for video stream from the camera via WIFI. As soon as the video stream reaches to processing unit, it starts analyzing the stream frame by frame in real time. Frames are then fed into object recognition and face recognition models to provide accurate results. Finally, the result is converted into voice signal using text to speech conversion technology and sent back to the bone conduction unit via Bluetooth. Then the wearer can listen the result as output.

If user input command is to search any information from the internet, user voice will be transferred to the processing unit through previously discussed process and processing unit will find related results from Wikipedia and speak to the bone conduction unit.

A custom designed workable eyeglass prototype shown in "Fig. 1(b)" has been printed in a 3D printer to fit all the hardware components and function the software features properly.

IV. Hardware Devices and Configuration

A. Camera

A Sony Exmor IMX219 Sensor based camera is used which has fixed focal length (2.96mm), 77.6-degree wide viewing angle and capable of capturing 1080Pixel 60 frames per second (FPS), 720Pixel 180 FPS, 8Mega Pixel pictures and videos.

B. Processing unit

Raspberry Pi has been used as the processing unit of the system because it is cheap, available and open sourced development board. It also supports different operating systems and computer vision libraries. It has built in WIFI and peripheral ports to connect other devices.

C. Battery and Charging Unit

1 lithium polymer 450mAh battery has been used to power up the system. When the battery is fully charged, it provides maximum of 4.2V output voltage. But most of the connected devices require constant 5V input. So, a MH-CD42 circuit is used which works as a boost converter to amplify and stabilize the battery output voltage. This circuit also works as charging and discharging unit of the system. "Fig. 2" represents the power management process diagram.

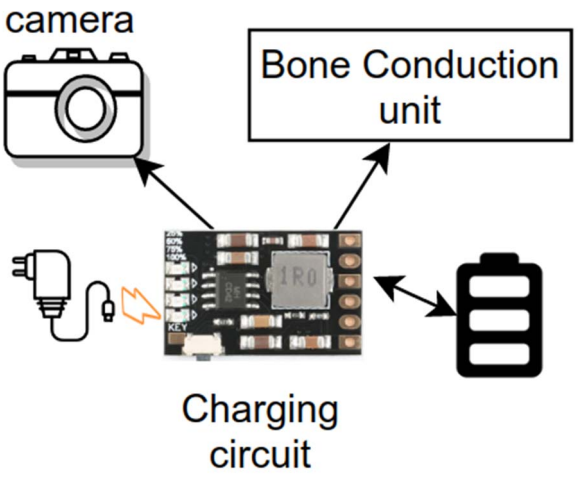


Fig. 2. Power Management Process of the system

D. Bone Conduction Unit

A customized bone conduction unit is used and embedded into the glass frame. Bone conduction technology is widely used in making hearing aid devices. It works by vibrating against the bones in the upper jaw and send the vibrations to the inner ear bypassing the ear canal completely. It gives freedom to the wearer to get the best outcome of eyeglass and listen to surrounding sounds simultaneously.

V. Software Configuration and Implementation Strategies

A. Voice Control and Conversion

Speech signal has been used as the input of the system. Currently the system supports English and Bengali Language. User provided speech is transferred to the processing unit via Bluetooth. Then the speech is converted intro machine readable text through an API call. Generated text is then used for further processing and processed output is converted to speech through another API call.

B. Assistance from Wikipedia

User can interact with the eyeglass and ask basic queries. If any result found for that specific query in Wikipedia, a response will be sent through bone conduction unit.

C. Object and Facial Recognition

Object and face recognition are two of the core features in this system. For both processes, flow chart showed in "Fig. 3" has been followed. User input given through the microphone is converted to text and compared it with the camera feed grabbing command. As soon as the camera feed is ready, it is sent for detecting objects and faces.

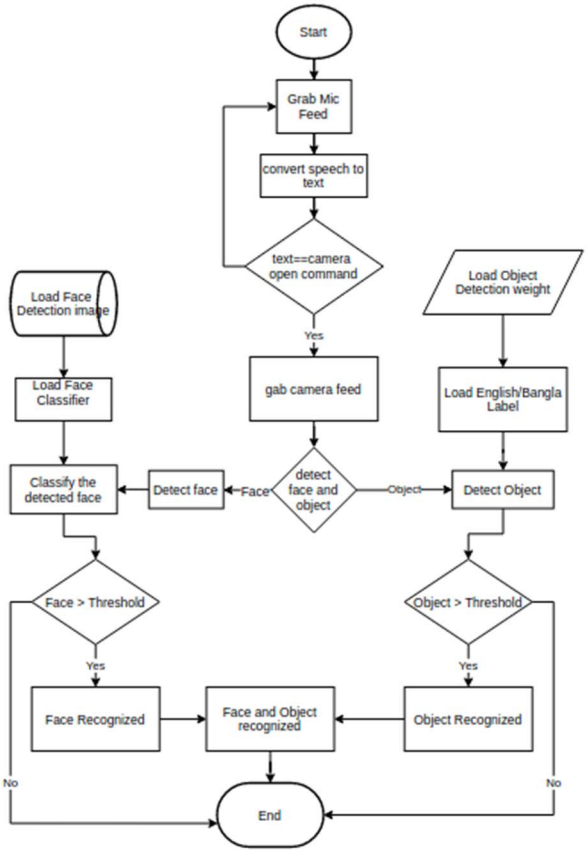


Fig. 3. Object and Face Recognition Flow Chart

1) Object Recognition

Image frames from the video feed are captured and implemented YOLOv3 [4] to detect objects. YOLOv3 predicts bounding boxes using dimension clusters, as anchor boxes likewise YOLO9000 [5]. for each box, it predicts 4 coordinates, x, y, height and width. YOLOv3 extracts object feature with a deep convolutional neural network called Darknet-53. Mini MS COCO pre-trained model weight is used in this implementation, which is trained on millions of images from different classes. Class label names have been customized in both languages and calculated model accuracy. If the model prediction is more than a certain threshold value, label name and corresponding accuracy are selected as output from the object recognition process.

2) Facial Recognition

MTCNN [6] model is used to extract faces from image frames. MTCNN is a three stage cascade framework. The stages are Proposal Network (PNet), Refine Network (R-Net) and the O - Net. MTCNN detects faces using five facial Landmark positions. After detecting faces a classifier is used to classify faces. For experimental purpose, up to 300 classes of facial images are tested and found satisfactory accuracy for each class. Support vector machine is implemented to classify detected faces from MTCNN output images. If the output accuracy of SVM is greater than a certain threshold, corresponding label name and classification accuracy is selected as final result from facial recognition process.

VI. Model Testing and Result Analysis

The eyewear is tested from different viewpoints to check the accuracy of simultaneously happening activities and collected feedbacks. It is mainly built for visual impaired people, so the device is taken out of experimental and development environment and given to a complete blind person who had lost eyesight after major injury. This is a good use case for testing the system usage sentiment from actual users. User feedback and experience have been collected to understand social impact of the device too.

As the entire system needs to fit in a wearable eyeglass, many additional parameters have been taken into consideration including total eyewear weight, power consumption and charging cycle, user experience and visual outlook. To keep total weight low but provide a satisfactory amount of active period, a 450mAh lithium polymer battery has been used. Battery performance has been observed for a month. The final charging and discharging cycle chart is given in figure "Fig. 4" and "Fig. 5" respectively. The prototype eyeglass weights 182 grams only which can also be reduced

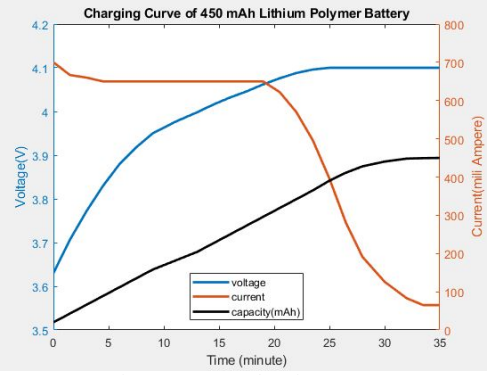


Fig. 4. Battery Charging Curve

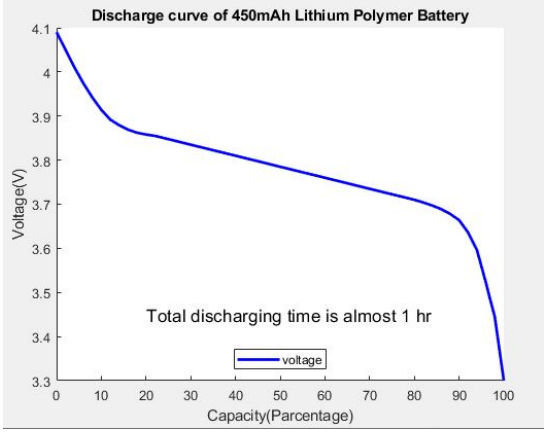


Fig. 5. Battery Discharge Curve

by using custom designed circuit and lightweight polymer frame.

The entire system fits in an eyeglass so it was challenging in prototype phase to completely run both models inside the eyeglass. To overcome the issue, pre-trained model weights are used for object and face recognition. Actual model training was performed in an Ubuntu 18.04 LTS, one of the popular Linux distribution, operating system based computer with Nvidia RTX 2070 graphics processing unit. Raspbian, a popular debian based OS for Raspberry Pi and IoT devices is used as operating system for the eyeglass processing unit.

Object detection model recognizes the objects from real time video feed with higher accuracy. In most cases it predicts object with 85% -100% accuracy which is very important for a blind or visually impaired person. "Fig. 6" and "Fig. 7" show the result of the object recognition model in both indoor and outdoor environments. The image frames have taken from the camera feed of the eyeglass. The model can detect most of the common indoor and outdoor objects.

In case of face detection, two different models have been tested to select the best performing model. First model name was Haar Cascade. It detects face based on five facial keypoints or Haar Features. Processing time of this model is faster than the second implemented model, MTCNN. But MTCNN performed better than Haar Cascade. It showed better accuracy in test images. To train and test the models, almost 8 thousand images of 15 persons had been captured within 2-month time window. All classes are distributed with equal number of images to reduce biases. Performance of both models is compared in table. 1.

Table 1. Face Recognition Models Comparison

Model name	Number of images	Number of detected faces	Recall score	Precision score
Haar Cascade	7628	6232	81.699%	91.527%
MTCNN Face detector	7628	7137	93.55%	96.832%

Fig. 6. Object Recognition from camera feed (indoor environment)

Fig. 7. Object Recognition from camera feed (outdoor environment)

As the device was designed for blind people, higher accuracy was prioritized over faster processing. Hence, MTCNN is used for face detection. Detected and extracted faces from image using MTCNN is shown in "Fig. 8". Actual names of the persons are renamed for anonymity and used generic names as class labels. Finally, "Fig. 9" represents the face classification result from a video feed in indoor environment.

Fig. 8. Faces detected and cropped by MTCNN model

ACKNOWLEDGMENT

The authors would like to thank Md. Sadique Sarwar from University of Dhaka and Md. Ehsanul Haque for sharing their ideas, implementation strategies and continuous support. Authors also thank the employees of Arollo Tech Limited who have voluntarily participated and consented to use their facial image data for model training, testing and publishing publicly.

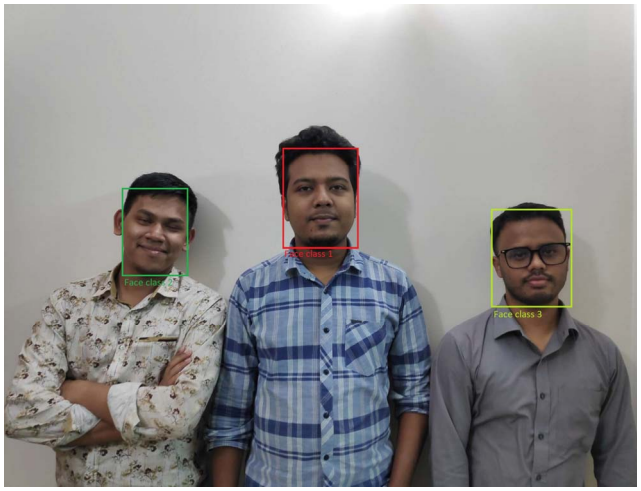


Fig. 9. Face Recognition from eyeglass Camera feed.

CONCLUSION

The proposed system represents a vision aid device in a form of eyewear for the blind and partially blind people with the help of modern computer vision technologies and computer hardware. The device detects objects, recognizes face, helps the wearer with information from Wikipedia, uses bone conduction technology for a better user experience. The authors believe that this device will help millions of blind people to decrease some of their daily life challenges and overcome the fear of unknowns.

REFERENCES

[1] Bourne RRA, Flaxman SR, Braithwaite T, Cicinelli MV, Das A, Jonas JB, et al.; Vision Loss Expert Group. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. Lancet Glob Health. 2017 Sep;5(9):e88897.

[2] Feng Lan, Guangtao Zhai, & Wei Lin. (2015). Lightweight smart glass system with audio aid for visually impaired people. TENCON 2015 2015 IEEE Region 10 Conference. doi:10.1109/tencon.2015.7372720.

[3] Vincent Ilardi, Renaissance Vision from Spectacles to Telescopes. Philadelphia, Pennsylvania: American Philosophical Society, 2007.

[4] Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. ArXiv, abs/1804.02767.

[5] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 65176525. IEEE, 2017.

[6] K. Zhang and Z. Zhang and Z. Li and Y. Qiao, Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Processing Letters. 2016, vol 23, page: 1499-1503.

[7] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, Face detection without bells and whistles, in European Conference on Computer Vision, 2014, pp. 720-735.

[8] S. Yang, P. Luo, C. C. Loy, and X. Tang, From facial parts responses to face detection: A deep learning approach, in IEEE International Conference on Computer Vision, 2015, pp. 3676-3684.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770778, 2016.

[10] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards realtime object detection with region proposal networks. arXiv preprint arXiv:1506.01497, 2015.