Image/Video Summarization in Text/Speech for Visually Impaired People

Chaitra C[1], Chennamma[1], Vethanayagi R[1], Manoj Kumar M V[1], Prashanth B S[1],
Snehah H R[1], Likewin Thomas[3], Shiva Darshan S L[2]
[1]*Department of Information Science & Engineering,* NITTE-Meenakshi, Bangalore 560064, Karnataka, India
[3]*Department of Information and Communication Technology,* Manipal Institute of Technology, Bengaluru-560 064, Karnataka, India
[3]*Department of Information Science & Engineering,*
PESITM, Shimoga 577204, Karnataka, India,
chennammaa11@gmail.com, 1nt19is402.chaitra@nmit.ac.in,
1nt19is415.vethanayagi@nmit.ac.in,{manojmv24, prashanth.bshivanna, likewinthomas, darshkottur}@gmail.com

2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon) | 978-1-6654-9790-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/MysuruCon55714.2022.9972653

***Abstract*—In the year 2022, an estimated 2.2 billion people around the globe will have a visual impairment. The problem may be hereditary or due to accidents. Nonetheless, technological advancements in helping visually impaired people have been going on for a long time. The admittance of technical concepts such as robotics, Machine Learning, and Artificial Intelligence for societal needs has proven worthwhile. The blind or visually impaired people learn about their surroundings through other senses, such as touch, hearing, and smell. Our proposed work aims to build an end-to-end solution for visually impaired people to help them grasp the environment by summarizing the images or video streams with the help of Machine Learning paradigms. The proposed work uses a pre-trained Caffe Object Detection model and requires less data for training and detection. We have developed a Client-Server model for our proposed idea wherein the significant computations happen on the server side, which is the Object detection model, and the client App is developed using Android. The app also has a text-to-signal processing feature that helps summarize the objects detected in the form of an audio catalog.**

***Keywords*—Visually Impaired, Caffe, Object Detection, Android**

I. Introduction

Image processing technology applies transformations to digital images to observe their valuable features. Intelligent and expert systems are getting increasingly integrated into our daily lives regularly. For contemporary technological advancements, creating intelligent systems to enhance the quality of life is of the utmost importance. Algorithms for machine learning and artificial intelligence are crucial elements to achieving the advancements [1] [2]. Many tools and programs are available to help the visually impaired in their daily life. Among the gadgets are mobility assistance robots, SCADA systems, sticks, optical sensors such as RGB cameras, and GPS. One of the issues with these programs is that they are created for a specific function rather than recognizing general objects in your environment. To process the objects, some applications require a live connection to the server or cloud [3]. The application dedicated to helping impaired people is usually made with the help of Microcontrollers and Mechanical equipment combined with the Internet of Things. Though these solutions are acceptable in terms of usage, they fail to be cost-effective when brought into production. This pitfall had been a research avenue to explore where contemporary approaches cascade multiple domains are used. Image processing is one of the most growing fields in research and technology in today's world [4]. When used in conjunction with Image processing paradigms, Machine Learning proved to be effective in providing competing solutions to the mechanical and IoT solutions. Our objective is to build an effective and affordable solution for the visually impaired so that they can visualize the environment with the help of handheld devices such as Mobile Phones. The application is designed for Android Smartphones and uses MobileNet as Neural Network for learning in the back-end server. The neural network categorization is shown in Fig. 1.

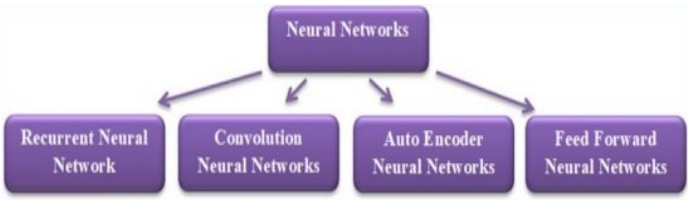

Figure 1. Neural Network Categories [5]

Recurrent Neural Networks(RNNs): The following information, namely time-stamped information from a sensing device or a spoken utterance, is determined progressively. The inputs to recurrent neural networks need not rely on one another, in contrast to classic neural networks; It depends on calculations made by its other components. to produce the desired results. Applications for forecasting and time series, sentiment analysis, and other text functions are applications utilizing RNNs.

Convolution Neural Networks(CNNs): This contains input, convolution, pooling, completely linked, and output Layers, among other Layer types. Every layer has a certain area of

Authorized licensed use limited to: University of Texas at Arlington. Downloaded on December 09,2023 at 07:14:06 UTC from IEEE Xplore. Restrictions apply.

expertise, such as connecting, activating, or reviewing. These CNNs can also be employed in various research projects, including forecasting and natural language processing.

Autoencoder Neural Networks(ANNs): These are created by some preoccupations, specifically encoders, which are created by a collection of provided inputs. Autoencoders typically evaluate themselves using an unsupervised paradigm in which they are compared to other conventional neural networks. Encoders are being used to transform senseless input into irrelevant information and social democrats data into relevant information following the model. Given that this model is tiered, higher layers may provide generalization when a decoder is enabled to the nearest point. These assumptions may be extensions or categories that are linear or nonlinear.

Feedforward Neural Networks: Along with the connections to the other layers, this network that connects the layers sequentially one after the other. This network's data is unidirectional, meaning that it can only be transmitted one way—progressively from one node to another. There won't be any feedback loops of any kind.

Due to their size and increased computational complexity, network designs like the one described above are challenging and complex in restricted devices. We use Mobilenet, which may be used on limited devices like smartphones, to solve this issue [6]. MobileNet's simplified architecture includes deep neural networks that can be built using depth-wise separable convolutions. Convolutional Networks can be distinguished from conventional ones. There are two stages to the general concept of depth-wise separable convolution: Convolution of 33 depths and Convolution of 11 points.

Following is the arrangement of the sections of this paper: The second section of this paper discusses the literature review of the existing approaches to Image or Video Summarization. The research problem is described in Section III. Section IV discusses the framework and system design. Section V provides details on the algorithms used in this work. Section VI describes the Collected empirical findings. The emphasis in Section VIII is on the conclusions and future scope.

II. Literature Review

Web-based applications are built to protect user privacy. The user of this application can switch on the on-demand feature to sacrifice privacy while updating the family on his or her status. They may recognize items around them and comprehend their immediate environment using a low-power Mobile-Net Architecture based on CNN. The experiment's participants had problems identifying road curbs, changes in the road's surface, and staircases when it was being carried out outside. The benefits and limitations of each type of soft fabric in inside situations, such as soft cushions and drapes, are also discussed [7].

In [8] an assessment of assistive technology using scores to illustrate how well they can add features, VIP assistive devices are categorized according to their features and operational principles. A quantitative study based on scores is performed to determine these devices' effectiveness and the possibility of feature augmentation for each category.

The idea of expanding the number of items that can be tallied at once was studied in [9] [10] in order to improve the support given to persons who are visually impaired. A voice alarm that can recognize all probable daily multiple objects will also alert consumers to both nearby and far-off objects. The earlier techniques are subject to limitations in terms of accuracy, scene complexity, illumination, and other factors.

By providing high precision, the best performance results, and accessible options, the objective is to improve the quality of life for blind or visually impaired people. The application's User Interface is self-explanatory and easy to use, making it appropriate for blind users [11].

The main goal is to assist a person who is blind indoors by just hearing an audio query about the required object and providing aural feedback, such as the name, distance, and direction of that particular thing. The best depth quality per depth and good accuracy enable a blind person to quickly calculate distance and direction. The results of the tests also show that this technology is easily able to overcome hurdles because it can move around in environments with few limits [12].

The model's computational efficiency has been enhanced with the help of the architecture of a faster region convolutional neural network. They removed feature data from the image and used feature mapping to assign a class label. The SSD algorithm's default boxes used with a multi-scale feature map mitigated this reduction [5].

The Stanford model, which delivers more in-depth information, is utilized to create captions once the visual elements of images are extracted using VGG16. First, picture visual features are extracted using a pre-trained VGG16 model. The Stanford model is then used to create a caption for them [13].

Researchers have developed walking aids that ensure the mobility and safety of people with visual impairments. Any framework for helping people with visual impairments should take into account the study as a strong foundation for a complete description of the essential elements. The sensors that go into making walking aids may not be used appropriately without undesirable results [14].

Outcome of Literature Survey and research gaps

Many visually impaired people around us are facing numerous challenges. One of the most common and complicated difficulties is when they have to navigate themselves in an unfamiliar environment. After reviewing the literature and conducting a survey, we decided to create an object detection summarization for visually impaired people. So that they can visualise a summary of the objects around them in their native language (braille). In addition, voice modulation is being used to recognize the objects. As a result, object detection applications using machine learning techniques are employed to aid or assist them.

Authorized licensed use limited to: University of Texas at Arlington. Downloaded on December 09,2023 at 07:14:06 UTC from IEEE Xplore. Restrictions apply.

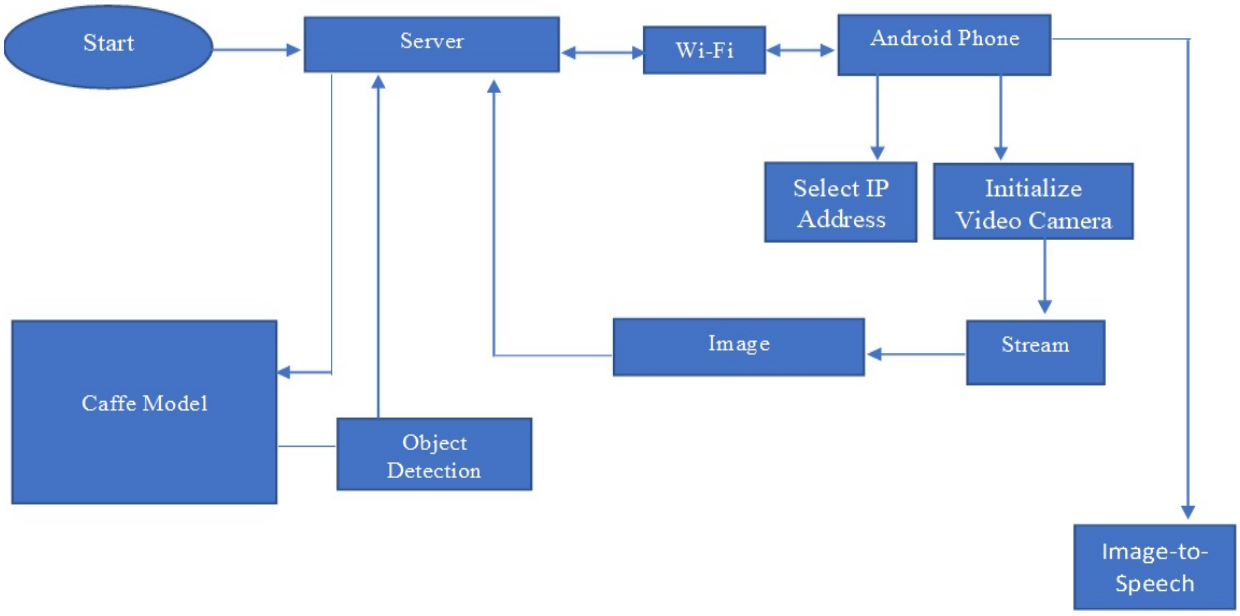

Figure 2. System Design of the Proposed Model

III. Research Problem

Recent technological developments have focused on creating practical systems to improve living standards. In order to help visually impaired people recognize objects, we can use machine learning. These people use their senses to learn about their surroundings by touching, hearing, and smelling, but they are limited in how much of their environment they can comprehend because they cannot use their eyes. Therefore, we decided to design an application called Image/video summarization in text/speech for visually impaired people to assist them so that they can record images of their surroundings and identify objects in them.

A. Contribtions of this paper

With gathering much information in the Existing system, many electronic devices like walking sticks and alternative devices were created to assist blind people. However, the disadvantage of these systems is that they use hardware for development which is expensive and can't be afforded by all. This hardware is heavier for blind people to carry. This drawback motivated the proposal of an affordable and portable system to find objects and narrate for the visually impaired and assist them in their surroundings. The main Objectives are:

- Smartphones have sensors integrated with cameras and speakers, thus reducing the development cost of application.
- Server and portable android application to detect the objects.
- Evaluate the accuracy of the proposed method with the existing methods.

IV. Framework and System Design

System design is designing a system's architecture, parts, and interfaces to satisfy the users' needs. Fig. 2 shows the proposed model's various phases, and each of their functionality is described below.

- Server: In this phase, after starting Computer. The server is used for sharing the information with other computers to detect the image and deliver the output to Android via wifi. The objects detected is from the trained set of the caffe model.
- wifi: Image from the server is being detected by wifi.
- Android Phone: After being detected by the server, this android device detects the image and voice.
- Select IP Address: In this phase, the android phone selects the IP Address, and the IP address of the phone and Computer should be the same.
- Initialize Video Camera: As soon as, the android phone and Computer is connected to the same IP address, the video camera can be initialized.
- Caffe Model: It is a machine learning model with an image classification or image segmentation model trained using caffe.
- Object Detection: The objects trained from the caffe model can be detected and sent to the server.
- Image: The objects which are trained from the caffe model are detected as soon as the video starts streaming.
- Stream: The video can be streamed as soon as the android phone is detected with the same IP Address.
- Image to speech: The object detected is converted to image as well as to speak.

Authorized licensed use limited to: University of Texas at Arlington. Downloaded on December 09,2023 at 07:14:06 UTC from IEEE Xplore. Restrictions apply.

V. Algorithm and Flow Chart

The Algorithm used in Object Detection is the Convulational Network Method (CNN). CNN has multiple layer's included.

- Convolution Layer: CNN has a convolution layer with several filters to perform the Convolution Layer.
- Rectified Linear Unit(ReLU): CNN has a ReLU layer that allows it to perform operations on elements. The output will be rectified in the future.
- Pooling Layer: The rectified feature map next to the pooling layer.

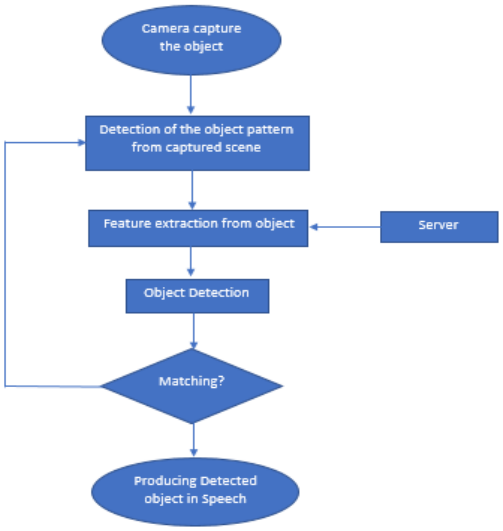

Figure 3. Stepwise implementation flow of the proposed model

$$L_loc(t^u, v) = \sum_{i \in x,y,w,h} smooth_L 1(t_i^u - v_i) \quad (1)$$

Back-propagation [15] through pooling layers of Region of Interest (RoI). Backpropagation is the process by which derivatives are routed through the RoI pooling layer. We assume only one image per mini-batch $(N = 1)$ for clarity, while the expansion to $N > 1$ is simple because the forward pass treats all images individually. 1442 collection of inputs in the sub-window that the output unit $y_r j$ pools the most of. A single x_i can be assigned to a number of separate $y_r j$ outputs.

$$\frac{\delta L}{\delta x_i} = \sum_r \sum_j \left| i = i^*(r,j) \right| \frac{\delta L}{\delta y_r j} \quad (2)$$

Here i is the index of an anchor in a mini-batch, and pi is the predicted probability of anchor i being an object [16].

A. Implementation

The proposed method's implementation entails carrying out activities to deliver outputs and monitor progress in relation to the work plan. Monitoring is the control of implementation to be on track and achieve the results.

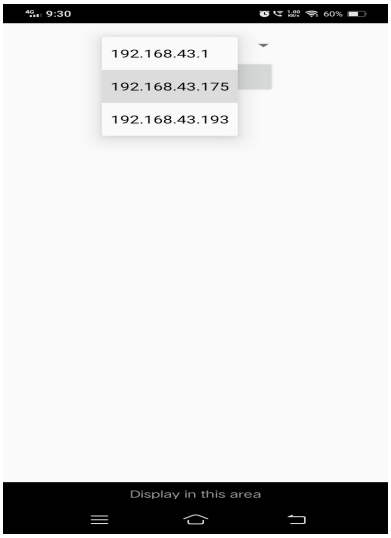

Figure 4. Shows how the IP addresses been detected and linked to wifi network.

B. Technologies Used

The main technologies that were used in implementing the system are Programming languages: java, python, and Android. Java is a high-level object-oriented language it is widely used for robust technology. Python is an interpreted and object-oriented language it has the ability to dynamic typic, exceptions, and other modules. Android – Android is used for creating an app. The trained model adopted is the MobileNet SSD caffe model.

C. List of Modules

The implementation of the modules is divided based on their operation and functionality. Here is the following list of modules

- Select IP: In this module, we detect all the IP addresses linked to the wifi network to which our android phone is connected. This list of IP addresses is added in a spinner, i.e., Combobox, so that the user can select the IP of the machine on which our object detection programming will be running.
- Start streaming: In this module, we capture live video from the camera and get frames, convert them into jpeg images, and send them to the Computer for object detection.
- Stop streaming: Here, video streaming is stopped, and the app gets disconnected to the connected PC.
- Switch camera: In this module, we can switch from the back camera to the front camera and vice versa in the android phone to capture live data from both cameras.
- Network Connection: In this module, we implement socket programming to connect from phone to the Computer for transferring images and receiving object names.

Authorized licensed use limited to: University of Texas at Arlington. Downloaded on December 09,2023 at 07:14:06 UTC from IEEE Xplore. Restrictions apply.

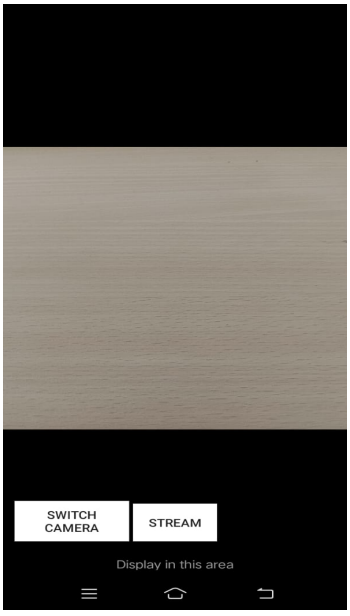

Figure 5. Shows two buttons, one for switching camera and the other for video streaming

Figure 7. Detects the object Bottle with text ”Bottle”

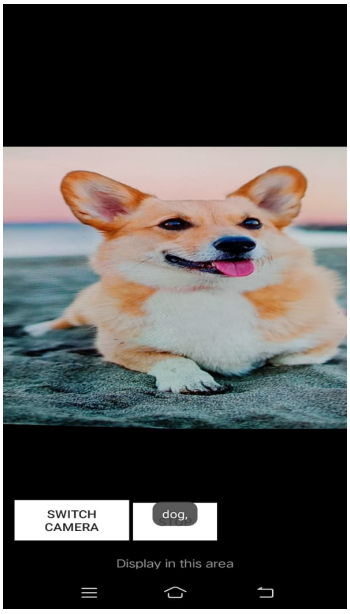

Figure 6. Detects the object Dog with text ”Dog”

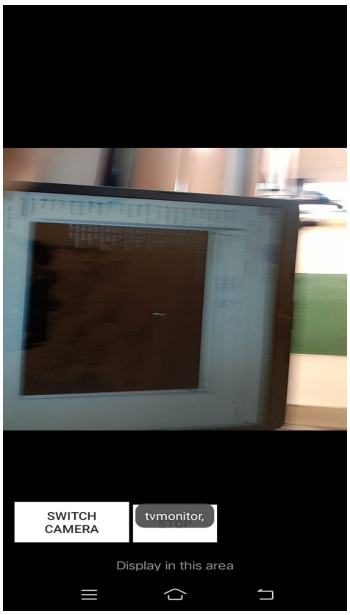

Figure 8. Detects the object Monitor with the text ”Monitor”

VI. Results and Discussion

The Fig. 4 searches all IP addresses connecting to the same network. Once the IP address is found, it connects with the server. The next result shows that the video streaming option, which has both front and back cameras, once the images are captured from the camera that will be converted into 300*300 pixels and rotates 270 degrees from the camera and gets frames converted to video format and sends frames to server if sent object matches with a trained object, server will give accurate results of objects.

After selecting the IP address, it has a video button by pressing video, it has two options one is to switch camera, and another one is streaming by clicking switch camera, we can switch back or front camera. The back camera gives more accuracy than the front camera and the streaming button is used for live capturing of images/video to detect objects.

Once we send an image to caffe model through the server, it has identified the object as a dog with Accuracy. If the Accuracy is greater than 0.5, then objects have been detected

Authorized licensed use limited to: University of Texas at Arlington. Downloaded on December 09,2023 at 07:14:06 UTC from IEEE Xplore. Restrictions apply.

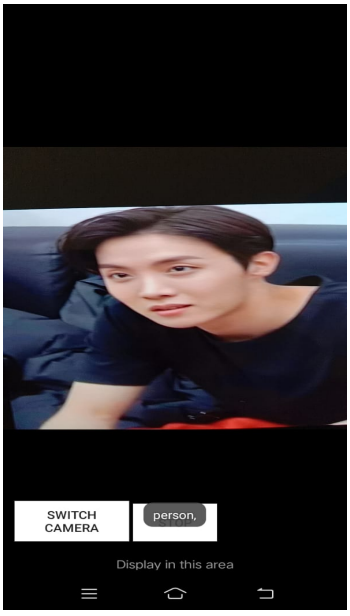

Figure 9. Detects the Person despite of gender with the text "Person".

successfully. If accuracy is lesser than 0.5, then results are not Accurate, as illustrated in figure 6. Similarly, the proposed model can detect other objects such as the bottle, monitor, and human, as shown in Fig. 7, Fig. 8, and 9 respectively.

VII. Conclusion and Future Scope

We put forth a system that can be utilized to help blind person comprehend their surroundings by describing the nearby things. Our proposed model will undoubtedly benefit visually impaired people. More emphasis has been placed on object detection and converting images to text and speech, ultimately aiding in object recognition. As an outcome, the above information gives clarity of what's been attained in assisting the visually impaired.

The life quality of blind persons is improved by image/video summarization applications so that they can replace darkness. The proposed work's future focus will be on recognizing several items in a view more accurately and quickly. Any sort of thing having a high frame rate can be characterized by this system's extension. The text-to-speech module was likewise created at a forward-thinking clip. Self-trained models can be used in place of pre-trained models. The model can be trained to visual features that the user sees regularly. As a result, it may be tailored to the user's unique demands and assures safer navigation. With the addition of a facial recognition capability, the application may be trained to remember details about those most intimately associated with the user, making it easier for them to distinguish between friends and strangers.

References

[1] Wenhao Chai and Gaoang Wang. Deep vision multimodal learning: Methodology, benchmark, and trend. *Applied Sciences*, 12(13):6588, 2022.

[2] BD Parameshachari, HT Panduranga, et al. Medical image encryption using scan technique and chaotic tent map system. In *Recent Advances in Artificial Intelligence and Data Engineering*, pages 181–193. Springer, 2022.

[3] Jawaid Nasreen, Warsi Arif, Asad Ali Shaikh, Yahya Muhammad, and Monaisha Abdullah. Object detection and narrator for visually impaired people. In *2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, pages 1–4. IEEE, 2019.

[4] Saeed Mian Qaisar, Raviha Khan, and Noofa Hammad. Scene to text conversion and pronunciation for visually impaired people. In *2019 Advances in Science and Engineering Technology International Conferences (ASET)*, pages 1–4. IEEE, 2019.

[5] Venkata Naresh Mandhala, D Bhattacharyya, B Vamsi, and N Thirupathi Rao. Object detection using machine learning for visually impaired people. *International Journal of Current Research and Review*, 12(20):157–167, 2020.

[6] Jing Zhang and Dacheng Tao. Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things. *IEEE Internet of Things Journal*, 8(10):7789–7817, 2020.

[7] Fahad Ashiq, Muhammad Asif, Maaz Bin Ahmad, Sadia Zafar, Khalid Masood, Toqeer Mahmood, Muhammad Tariq Mahmood, and Ik Hyun Lee. Cnn-based object recognition and tracking system to assist visually impaired people. *IEEE Access*, 10:14819–14834, 2022.

[8] Sadia Zafar, Muhammad Asif, Maaz Bin Ahmad, Taher M Ghazal, Tauqeer Faiz, Munir Ahmad, and Muhammad Adnan Khan. Assistive devices analysis for visually impaired persons: a review on taxonomy. *IEEE Access*, 2022.

[9] Mansi Mahendru and Sanjay Kumar Dubey. Real time object detection with audio feedback using yolo vs. yolo_v3. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 734–740. IEEE, 2021.

[10] Wenwu Zhu, Xin Wang, and Wen Gao. Multimedia intelligence: When multimedia meets artificial intelligence. *IEEE Transactions on Multimedia*, 22(7):1823–1835, 2020.

[11] Sunit Vaidya, Naisha Shah, Niti Shah, and Radha Shankarmani. Real-time object detection for visually challenged people. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 311–316. IEEE, 2020.

[12] Syed Sammak Hussain, Dua Durrani, Abdul Aziz Khan, Resham Atta, and Lubaid Ahmed. In-door obstacle detection and avoidance system for visually impaired people. In *2020 IEEE Global Humanitarian Technology Conference (GHTC)*, pages 1–7. IEEE, 2020.

[13] Burak Makav and Volkan Kılıç. A new image captioning approach for visually impaired people. In *2019 11th International Conference on Electrical and Electronics Engineering (ELECO)*, pages 945–949. IEEE, 2019.

[14] Md Milon Islam, Muhammad Sheikh Sadi, Kamal Z Zamli, and Md Manjur Ahmed. Developing walking assistants for visually impaired people: A review. *IEEE Sensors Journal*, 19(8):2814–2828, 2019.

[15] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Authorized licensed use limited to: University of Texas at Arlington. Downloaded on December 09,2023 at 07:14:06 UTC from IEEE Xplore. Restrictions apply.