International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 Impact Factor: 8.176 ISSN: 2582-3930

Netraa: A frame descriptor for visually challenged

Dikshita Padte[1],Karunesh Palekar [2],Ekta Vayanan[3], Prof. Deepti Vijay Chandran[4]
[1]*Student, Computer Engineering, SIGCE, Navi Mumbai, Maharashtra, India*
[2]*Student, Computer Engineering, SIGCE, Navi Mumbai, Maharashtra, India*
[3]*Student, Computer Engineering, SIGCE, Navi Mumbai, Maharashtra, India*
[4]*Asst. Professor, Computer Engineering, Smt. Indira Gandhi College of Engineering, Navi Mumbai, Maharashtra, India*

Abstract-

Visually challenged people face a lot of problems in their day-to-day life making their life more challenging than others. For this, there is a need for various technologies which could help them to overcome these difficulties in their day-to-day life. For this, we have surveyed some of the applications and technologies used which could be the eyes of visually challenged people enabling them to visualize the surrounding, use the smartphone and carry out all the activities which a person with vision does. In this project, we have implemented EfficientDet Object Detection Model to help visuallyimpaired people to visualize their surroundings, detect the objects in front of them and get to know the obstacles in their path beforehand using speech.

Key Words: Visually Challenged, Frame Descriptor, Efficient Det, TensorFlow Lite, Text to speech, Netraa

1. INTRODUCTION

Dealing with sight loss, already, is a challenge in itself. The lack of emotional support at diagnosis centers, the limited accessibility to activities and information, the societal stigma, and the lack of unemployment, are all factors frequently leading blind or low-vision individuals to isolation. Work is a whole different matter if you're visually impaired. Considering the lack of accessible work and working spaces, one can already imagine why hiring a visually impaired individual would beconsidered a liability for a company. This harms the confidence and emotional well-being of the visually impaired, while it cripples their economic independence. Having little to no opportunity to support themselves, blind or low-vision individuals are incapacitated from theirindependence. Hence there is a need to come up with such a technology that could help them to gain equal respect and opportunities in the workspace and also in the outdoor world.

2. LITERATURE SURVEY

2.1 Survey of Existing System

It's estimated that there are about 36 million people in the

world who are blind, and a further 216 million who live with moderate to severe visual impairments. Although the World Health Organization points out that up to 80% of vision impairment around the world is avoidable with better access to treatment, the number of people who are blind or have low vision is rising as theglobal population ages. But technology is playing a vital role in tearing down barriers, and artificial intelligence is making real inroads into improving accessibility.
Microsoft's Seeing AI is an app designed to help people with low vision or who are blind. It enhances the world around the user with rich audio descriptions. It can read a handwritten noteor scan a barcode and then tell the user what the product is. Point a camera at something and the app will describe how many people it can see and where they are in the image – center, top leftand so on.

2.2 Research Paper Analysis

G.Lavanya ME., Preethy. W, Shameem.A and Sushmitha. R, "Passenger BUS Alert System for Easy Navigation of Blind", Int. Conf. on Circuits, Power and Computing Technologies [ICCPCT-2013], 2013, pp.798-802[1] In this paper, a Smart Assistive Navigation System for Blind and Visually Impaired Individuals is designed and implemented to secure safe and low-cost navigation. Blind navigation systems many, but very few are those that are completely successful in addressing the requirements of blind individuals to navigate safely, comfortably, and independently. Thus, some projects of state-of-the-art are discussed and analyzed. Finally, the design and the implementation of the system are shown**. HE, Kaiming, ZHANG, Xiangyu, REN, Shaoqing, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770-778.[2]**Indoor signage plays an essential component to find destinations for blind and visually impaired people. In this paper, we propose an indoor signage and door detection system to help blind and partially sighted persons access unfamiliar indoor environments. Our indoor signage and door recognizer is built based on deep learning algorithms. We developed an indoor signage detection system specially used for detecting four types of signage: exit, wc, disabled exit, and confidence zone. Experiment results demonstrate the effectiveness and the high

precision of the proposed recognition system. We obtained 99.8% as a recognition rate. **T. Kapic. Indoor navigation for visually impaired." www.mics.ch/SumIntU03/TKapic.pdf, 2003.[3]**A wearable navigation system for visually impaired and blind people in unknown indoor and outdoor environments is presented. This system will map and track the position of the pedestrian during the exploration of the unknown environment. To build this system the well-known Simultaneous Localization and Mapping (SLAM) from mobile robotics will be implemented. Once a map is created the user can be guided efficiently by a route-selecting method. The user will be equipped with a short-range laser, an inertial measurement unit (IMU), a wearable computer for data processing, and audio-bone headphones. This system does not intend to replace the use of the white cane. However, the purpose is to gather contextual information to aid the user in navigating with the white cane. **Dąbrowski, P. Kardyś, T. Marciniak, "Bluetooth technology applications dedicated to supporting blind and hearing as well as speech handicapped people", The 47th IEEE ELMAR Symp., Zadar 2005, pp. 295–298.[4]**This article describes a new Android application supporting blind and partially sighted people in smartphone use. It enables them to call, send and receive text messages, make use of a "phone book" as well as of additional options such as positioning or battery monitoring, through voice commands. The software concept together with the structure of the respective application has been presented in detail.**[6]S. K. Chaitrali, A. D. Yogita, K. K. Snehal, D. D. Swati and V. D. Aarti, "An Intelligent Walking Stick for the Blind", International Journal of Engineering Research and General Science, vol. 3, no. 1, pp. 1057-1062, 2015.[5]**This paper presents the idea of developing a smart system that can assist visually impaired people in their daily activities. There are many challenges faced by visually impaired people. In most cases, they require constant support in almost all scenarios, especially in their day-to-day activities. Some of the major challenges include difficulty in moving from one place to another without the assistance of someone. Other challenges include difficulty in recognizing people, detecting obstacles, etc. To counter avert this situation, we propose a "smart eye system" in this work. The device is a voice-enabled system that would direct the visually challenged person in their day-to-day work. The device combines the various available technologies and integrates them into a single multipurpose device that can be used by the visually impaired. The paper discusses the design of such a system and the challenges involved in designing the device.

2.3 Limitation on Existing system

Many applications, software, and systems have been already made to assist in either way. But no software is so capable to include all the technologies which could help the visually challenged person in every sense. Visually challenged people cannot completely rely on these existing apps and models for living. The existing models are capable of performing only object detection but none of them can describe the environment so accurately using speech. Hence we tend to combine all the existing and more innovative technologies in this project.

2.4 Limitations faced during the implementation of the project

- Data Collection is a tough task here, as a diverse type of data has to be used for model training to get high accuracy.
- As the dataset used here is not extremely large, the model does not exhibit good accuracy.
- Also, frame description is a tough task in itself as the outputs of individual models have to be combined to obtain a description of the frame which includes background information, main object, and action.
- Model generalization is not up to the mark.
- Hardware limitation: Google Colab offers less memory required for model training. Also, GPU processing is another important issue, as sometimes the GPU speed offered by Google Colab does not support model training.

3. PROBLEM STATEMENT

By taking into consideration the problems of visually challenged people, this project has been implemented using the EfficientDet Object Detection model where the Object Detection Model is used to identify the day-to-day objects and obstacles coming in front of the visually challenged person. Not only objects but we used Efficient Det for posture detection, action detection, background detection, and so on. It further helps to detect currency and text reading. Every information detected by the model would be converted into voice snippets for communication of the information to the user audibly.

4. METHODOLOGY

Object detection will be performed on Images and video frames to identify objects such as chairs, tables, etc not only the objects would be detected but the complete frame description will be done to the user in the form of a speech snippet. For Eg. People are walking on the road, and A dog is running in the garden. Based on the images passed to the model, the model will be trained on a set of training images and further, the model will make a decision and perform object detection, Object recognition. Once detection and recognition are done, the image would be classified as Human or Animal using an Image Classification Algorithm. Further posture detection and background detection are accomplished and all these individual detections will be combined and an integrated output would be given. Further, this would be converted into a sentence to be given as output in the form of speech called a Frame description using Text to speech API... After the model is trained and tested it would be deployed using TensorFlow Lite and integrated with Android to make it compatible with Android Phones. The speech output is given with the help of Android Speech APIs. Once the Frame description is accomplished accurately, color identification, path awareness, and emotion detection can also be done.

5. PROPOSED SYSTEM

5.1 Data Collection from Google using a bulk image downloader

The Data was collected from Google using Fatkun Image Downloader. Here three parameters were taken into consideration(Background, Main Object, and Action). Initially, only Human and animal data were considered for simplicity.

5.2 Data Processing

In data pre-processing, there are steps involved such as image orientations, resizing, contrasting, data annotation, and data augmentations. Data annotation is a crucial part as the image has

to be labeled and a bounding box has to be drawn for training the actual Object detection model. Further, as the Android app would be fed with a real-time video frame, Data Augmentation was important to make sure that the app would process the image even if it is clicked from multiple angles.

5.3 Data Organization

- The data consist of Human and Animal Data. The human data set has three parts (Children, Adults, and Old age people) with male and female genders for each.
- Each cluster of Data(male and female both) has around 1500 images.
- In the children, the backgrounds chosen were parks, schools, and homes.
- In the adult cluster, the backgrounds chosen were office, home, and road.
- For old age, the backgrounds chosen were Home, park, and hospital.
- The data is then passed for annotation and Labeling augmentation.

5.4 EfficientDet Model: Everything you need to know and how EfficientDet works.

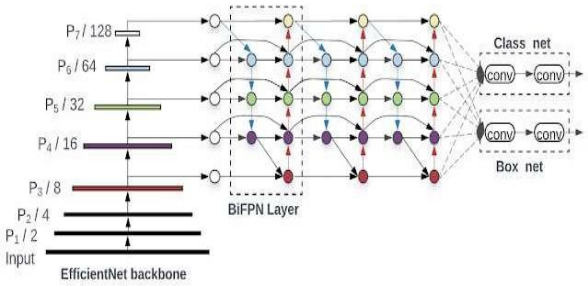


Fig 1: EfficientDet Model

Recently, the Google Brain team released their ConvNet model called EfficientNet. **EfficientNet forms the backbone of the EfficientDet** architecture, so we will cover its design before continuing to the contributions of EfficientDet. EfficientNet set out to study the scaling process of ConvNet architectures. There are many ways turns out - that you can add more parameters to a ConvNet.

EfficientNet set out to define an automatic procedure for scaling ConvNet model architectures. The paper seeks to optimize downstream performance given free rein over depth, width, and resolution while staying within the constraints of target memory and target FLOPs. They find that their scaling methodology improves the optimizationof previous ConvNets as well as their EfficientNet architecture.

Model efficiency has become increasingly important in computer vision. First, we propose a weighted bi-directional feature pyramid network (BiFPN), which allows easy and fast multiscale feature fusion; Second, we propose a compound scaling method that uniformly scales the resolution, depth, and width for all backbone, feature network, and box/class prediction networksat the same time. Based on these optimizations and better backbones, we have developed a new family of object detectors, called EfficientDet, which consistently achieve much better efficiency than prior art across a wide spectrum of resource constraints. In particular, with a single model and single-scale, our EfficientDet-D7 achieves state-of-the-art 55.1 AP on COCO test-dev with 77M parameters.

5.5 The Squeeze-and-Excitation (SE) block

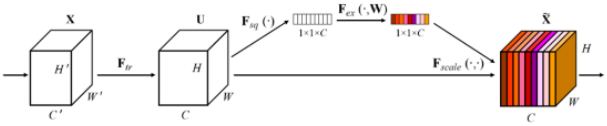


Fig. 1. A Squeeze-and-Excitation block.

Fig2: The Squeeze-and-Excitation (SE) block

The Squeeze-and-Excitation (SE) block is intended to improve the quality of a convolutional neural network's representations. For any layer of a convolutional neural network, we can build a corresponding SE block that recalibrates the feature maps:

- In the "squeeze" step, we use global average pooling to aggregate feature maps across their spatial dimensions H x W to produce a channel descriptor.
- In the "excitation" step, we apply fully-connected layers to the output of the "squeeze" step to produce a collection of per-channel weights ("activations") that are applied to the feature maps to generate the final output of the SE block.

5.6 Mobile V2 Backbone

In MobileNetV2, there are two types of blocks. One is residual block with the stride of 1. Another one is a block with a stride of 2 for downsizing.There are 3 layers for both types of blocks. This time, the **first layer** is **1×1 convolution with ReLU6.**

The **second layer** is the **depthwise convolution**.The **third layer** is another **1×1 convolution but without any non-linearity.** It is claimed that if ReLU is used again, the deep networks only have the power of a linear classifier on the non-zero volume part of the output domain.The main contribution is a novel layer module (the inverted residual with linear bottleneck), It takes as an input a low-dimensional compressed representation which is first expanded to high dimension and filtered with a lightweight depthwise convolution. Features are subsequently projected back to a low-dimensional representation with a linear convolution. Another change is to remove non-linearities in the narrow layers to maintain representational power.

5.7 Dynamic Convolution

Dynamic convolution [1] is a novel operator design that increases model complexity without changing the network depth or width by aggregating multiple kernels dynamically based on the attentions dependent on the input.
It is flexible enough to get placed into the existing CNN architectures and can further help to improve their accuracy.
By replacing the static convolution with dynamic convolution in MobileNetV2 and MobileNetV3, the method gains a top-1 accuracy improvement of 4.5% and 2.9% respectively on a 100M Multi-Adds budget with ~4% increase in the computational cost. For computing the kernel attentions, the squeeze-and-excitation method is used.First, the global average pooling layer squeezes the spatial information.

- Next, to generate the normalized attention weights for the convolution kernels, the input is further passed through two fully connected (FC) layers with a ReLU after the first layer and softmax after the second.
- Once the aggregated convolution completes after the attentions are computed, the output is passed through a batch normalization layer followed by an activation function (ReLU).
- The mentioned procedure builds a dynamic convolution layer.

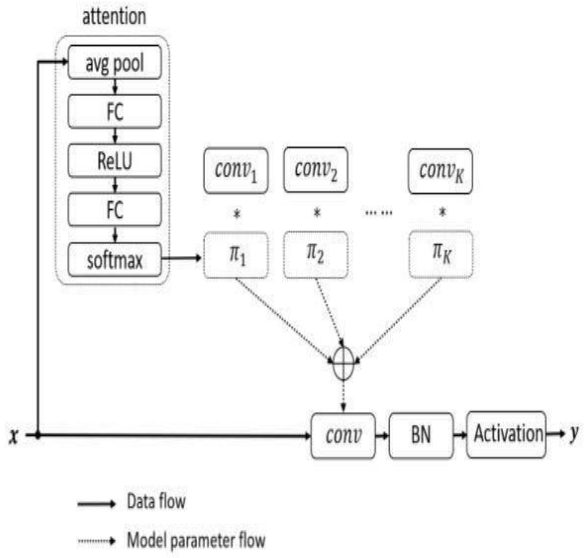


Figure 3: A dynamic convolution layer (Source: [1])

5.8 TensorFlow Lite Integration

TensorFlow Lite is a set of tools that enables on-device machine learning by helping developers run their models on mobile, embedded, and edge devices.
When one uses TensorFlow to implement and train a machine learning algorithm, one typically ends up with a model file that takes up a lot of storage space and needs a GPU to run inference. On most mobile devices, luxuries such as huge disk space and GPUs are not usable. TensorFlow Lite is a solution for running machine-learning models on mobile devicesThe TensorFlow Lite is a special feature and mainly designed for embedded devices like mobile.This uses a custom memory allocator for execution latency and minimum load. It is also explaining the new file format supported by Flat Buffers. TensorFlow Lite is a mobile library for deploying models on mobile, microcontrollers, and other edge devices.TensorFlow Lite takes existing models and converts them into an optimized version within the sort of .tflite file.

TensorFlow Lite algorithm

1. Step 1: Load Input Data Specific to an On-device ML App. The flower dataset contains 3670images belonging to 5 classes.
2. Step 2: Customize the TensorFlow Model. Create a custom image classifier model based onthe loaded data.
3. Step 3: Evaluate the Customized Model.
4. Step 4: Export to TensorFlow Lite Model.

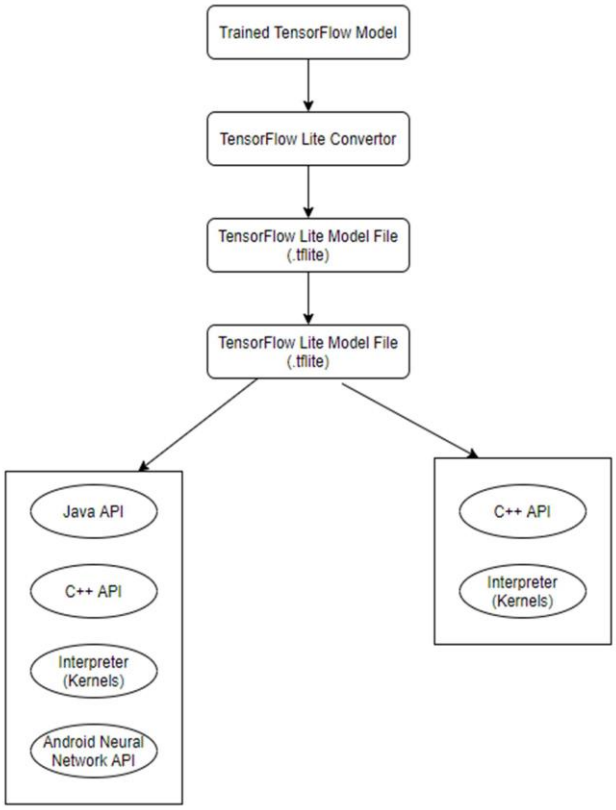


Fig 4: TensorFlow Lite

5.9 Text-to-Speech

Text-to-speech (TTS) is a type of assistive technology that reads digital text aloud. It's sometimes called "read-aloud" technology. TTS can take words on a computer or other digitaldevice and convert them into audio.
The voice in TTS is computer-generated, and reading speed can usually be sped up or sloweddown. Voice quality varies, but some voices sound human.

Steps for Converting Text to Speech in Android
Step 1: Create a New Project

Step 2: Working with the activity_main.xml file

Go to the app -> res -> layout -> activity_main.xml section and set the layout for the app. In thisfile add an EditText to input the text from the user, a Button, so whenever the user clicks on theButton then it's converted to speech, and a TextView to display the GeeksforGeeks text.
Step 3: Working with MainActivity.java fileStep 4: Output: Run on Emulator.

6. SYSTEM DESIGN

Design is a significant engineering illustration of whatever that's to be developed.

Program design is a process design that is an excellent option to effectively translate necessities into completed application products. The design creates a representation and presents the element of software information structure, architecture, interfaces, and add-ons which are vital to put into effect a procedure.

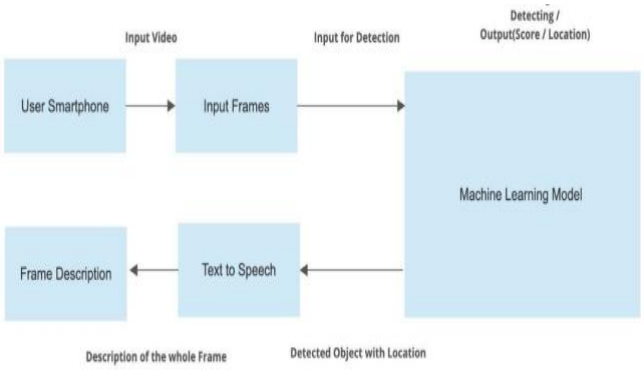


Fig 5: System Design

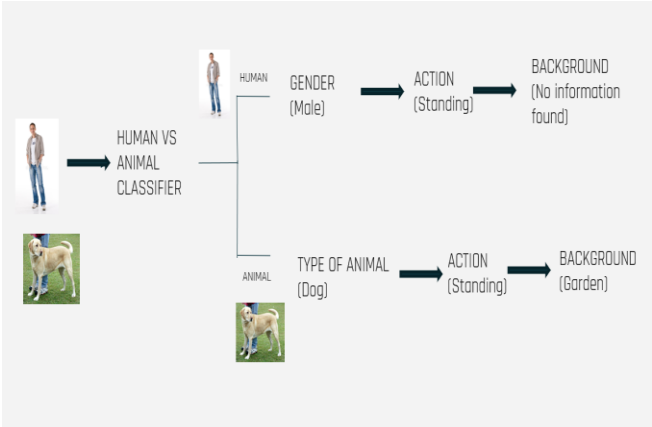


Fig 6: Image Classification Process

7. FLOWCHART

- A Data Set was created using Google images.
- The data consist of Human and Animal Data.
- The human data set has three parts (Children, Adults
- and Old age people) with male and female genders for each.
- The data is then passed for annotation
- and Labeling and augmentation.
- If the input data is
- a video is divided into frames and
- stored.

- The data is sent into an Image Classification model to check whether it is a human or an animal.
- The
- frames/images are further passed to the EfficientDet model for output where object detection and object recognition is performed.
- The output is categorized into various indoor and outdoor images which a person sees in daily life.
- The object, the action of the object, and the background are detected and recognized individually along with the action detection, and further, the individual outputs are integrated to obtain a combined result for frame description.
- The Model is deployed using TensorFlow Lite to use it on mobile devices.
- It undergoes Android integration.
- Further, all the output received from the model is converted to speech by using Text to speech API.

7.1 Flowchart for Data Collection

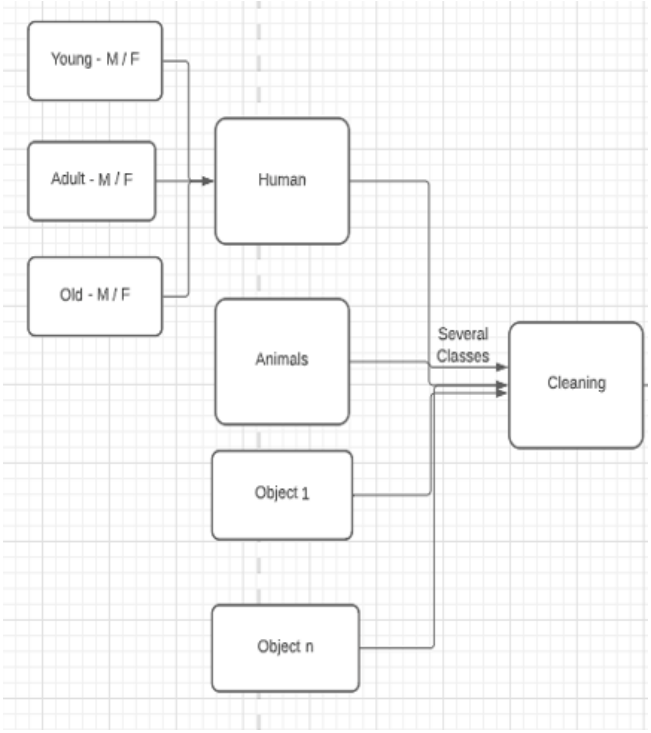


Fig 7: Flowchart for Data Collection

7.2 Flowchart for Data Preprocessing

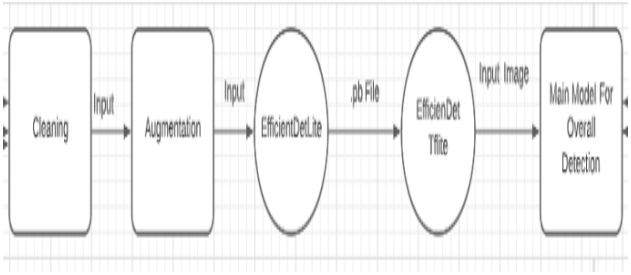


Fig 8: Flowchart for Data Processing

7.3 Flowchart for Model Implementation

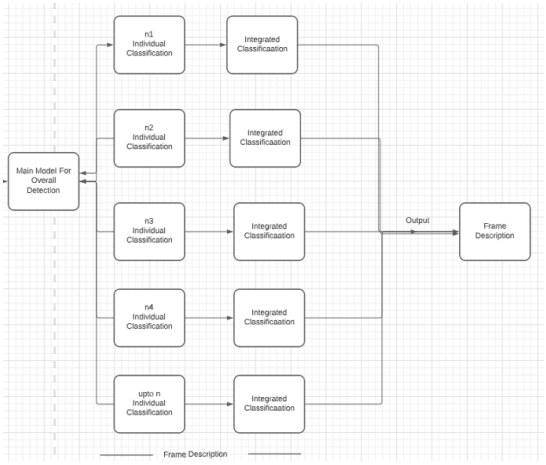
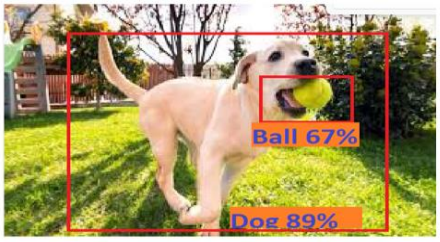


Fig 9: Flowchart for Model Implementation

8. RESULT ANALYSIS



A Dog is running in the Garden with Tennis Ball in the mouth.

Fig 10: Output

Special Features:

- **Torch**: When the mobile app is subjected to processing a real-time video frame, the visibility of the image or the object matters for processing the correct output accurately. Hence, the torch light features help to get clear input in low visibility areas. The torch of the mobile device will be automatically switched on under low visibility locations.
- **Focus**: The focus helps to highlight the man object in the video frame. This helps the model to take a clear input without any ambiguities.
- **Voice Recognition:** It recognizes the voice inputs given by the user using speech recognition and performs the tasks or gives answers accordingly.
- **Read the allocated text**: If a certain text is written on paper or through any digital device it can easily read it. This may help the visually challenged community to chat with their friends, and understand what's written in the message. More importantly, education can be made very easy for visually challenged children as they can easily understand what is written in the book by hearing through the app.

9. SCOPE

In a world where the next great invention is expected to appear on mobile phone screens, blind and visually impaired people have been left behind.Affordable smartphone apps have empowered the blind and the visually impaired. Now, **artificial intelligence (AI)** is taking those apps' capabilities to the next level. AI and machine learning technologies, specifically computer vision, have grown sufficiently robust to improve the lives ofthe blind and visually impaired. People with vision loss can do numerous things such as writing documents, browsing the internet, andsending and receiving emails. Screen Reading software and special talking and Braille devices allow those of us with no vision to use computers, cell phones, and other electronic devices independently.

This technology – commonly known as assistive or adaptive technology – is continually evolving and has removed many access barriers for people with vision loss. Besides allowing us to carry out routine tasks at work and school, assistive technology also enablespeople with visual impairments to be more independent at home. We can now read the mail, listento audiobooks, get step-by-step walking directions to unfamiliar places, record important information and so much more with special standalone devices designed for people with no or low vision.

10. CONCLUSION

With the help of this project, visually challenged people can detect objects or obstacles in their path without asking for human assistance. The app would give a detailed description of the obstacles like it's size, shape and distance. Education for the visually challenged can be made easier than before as the app will read the text inside the image fed to the system. This app can be used to enhance the safety of the visually challenged community, as it may alarm them about any obstacle or danger coming their way. Currency detection can also be performed which could help them not to get fooled by performing money exchange with vendors or other people. The voice recognition feature can cater to the specific requirements of the user and guide them whenever needed.

11. FUTURE ENHANCEMENT

This project is not complete in itself as it can be made more accurate by training with more massive datckgrounds but when trained with more data, this can also be made possible. Also, the model can be trained with the users' friends and family members so that they can correctly identify the person with their name while performing the detection.
The text reading feature can be made more paced which can help users to read novels or any other books easily.
Further facial recognition, braille, color identification, and distance of the obstacle can also be done as future enhancement in this project.

12. REFERENCES

- [1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition,2014.
- [2] R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), 2015.
- [3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, 2017.
- 4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
- [5] H. Caesar, J. Uijlings, and V. Ferrari, "COCO-Stuff: Thing and Stuff Classes in Context," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
- [6] "TensorFlow Lite | TensorFlow," TensorFlow. [Online]. Available: https://www.tensorflow.org/lite. [Accessed: 24-Mar-2019].
- [7] "EfficientDet | https://towardsdatascience.com/a-thorough-breakdown-of-efficientdet-for-object-detection-dc6a15788b73
- [8] G.Lavanya ME., Preethy. W, Shameem.A and Sushmitha. R, "Passenger BUS Alert System for Easy Navigation of Blind", Int. Conf. on Circuits, Power and Computing Technologies [ICCPCT-2013], 2013, pp.798-802
- [9] HE, Kaiming, ZHANG, Xiangyu, REN, Shaoqing, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770-778.
- [10] T. Kapic. Indoor navigation for visually impaired." www.mics.ch/SumIntU03/TKapic.pdf, 2003.
- [11] Dąbrowski, P. Kardyś, T. Marciniak, "Bluetooth technology applications dedicated to supporting blind and hearing as well as speech handicapped people", The 47th IEEE ELMAR Symp., Zadar 2005, pp. 295–298.
- [12] S. K. Chaitrali, A. D. Yogita, K. K. Snehal, D. D. Swati and V. D. Aarti, "An Intelligent Walking Stick for the Blind", International Journal of Engineering Research and General Science, vol. 3, no. 1, pp. 1057-1062, 2015.
- [13] Asad Ali Shaikh, Jawaid Nasreen, Warsi Arif, Asad Ali Shaikh, "Object Detection and Narrator for Visually Impaired People", 2019 6th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS)
- [14] Samir Patel, Amit Kumar," Smartphone-based Obstacle Detection for Visually Impaired People", 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS).
- [15] S. Usharani1, P. Manju Bala2, R. Balamurugan3," Voice-Based Form Filling System For Visually Challenged People", Voice-Based Form Filling System For Visually Challenged People.
- [16] S. Usharani1, P. Manju Bala2, R. Balamurugan3," Voice-Based Form Filling System For Visually Challenged People", Voice-Based Form Filling System For Visually Challenged People.
- [17]Aditya Dixit and VR Satpute," SIFRS: Spoof Invariant Facial Recognition System (A Helping Hand For Visual Impaired People)
- [18] Chen, Yinpeng, et al. "Dynamic convolution: Attention over convolution kernels." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020