Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 218 (2023) 1144–1155

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

# Third Eye: Object Recognition and Speech Generation for Visually Impaired

Koppala Guravaiah[a,*], Yarlagadda Sai Bhavadeesh[a], Peddi Shwejan[a], Allu Harsha Vardhan[a], Lavanya S[a]

[a]*Indian Institute of Information Technology Kottayam, Kerala, India - 686635*

**Abstract**

Detecting and recognizing the objects and generating speech about the objects helps visually impaired in a great way in understanding their surroundings. Required a mechanism to assist the visually impaired person to travel independently with the ability to identify objects in their path, and the ability to generate speech describing the objects detected in the scene. This can be achieved with the help of YOLOv5 image detection model and text to speech converters such as gTTS and pyttsx3 modules in python. The proposed method called Third Eye, giving better accuracy in detection and speech generation to help the visually impaired people Compared to existing approaches. YOLO v5 is trained on custom dataset of 15 objects along with MS COCO 2017 Dataset of 80 objects (95 objects overall). The output labels of the model are transformed to text and later converted to audio format and are presented to the visually impaired, through a speaker. We compared two python libraries for audio conversion, one is pyttsx3, and the other is gTTS.



*Keywords:* You Only Look Once (YOLO); Convolutional Neural Network (CNN); pyttsx3; gTTS; MS COCO; Text to Speech;

## 1. Introduction

Visually Impaired faces a lot of difficulties in their daily lives. There is a say that, Out of all the five sense organs, eyes are most important. The eyes are one of our most vital sense organs: 80% of what we perceive comes from our sense of sight [8]. According to World Health Organization (WHO), nearly 2.2 billion individuals have a close or faraway vision impairment. Out of them, 49.1 million individuals are visually impaired. Yet the growth of the population is making a substantial improvement in the number of people affected. There are significant inter-regional and gender disparities, highlighting the need to scale up vision impairment prevention programs at all levels.

* Corresponding author. Tel.: +91-970-386-3989.
*E-mail address:* kguravaiah@iiitkottayam.ac.in

1877-0509 
10.1016/j.procs.2023.01.093

The visually impaired always need the help of either a stick or a person. Early-onset severe vision impairment can restrict a child's verbal, emotional, social, and cognitive development, which can have long-term effects. Vision impairment critically impacts the quality of life among the adult population. Social isolation, difficulty in walking, a higher risk of falls and fractures, and premature admission to a nursing or home care are result due to vision impairment in people. As a result, proposed this work to assist visually impaired persons in recognizing their surroundings.

In recent years, deep learning has become a more popular technique for solving these kind of problems to identify the objects [11, 12, 29, 27, 14]. The deep learning systems achieve high accuracy rates at a lower cost. Many Convolutional Neural Network (CNN) methods such as Single Shot Detector (SSD) [13] and You Only Look Once (YOLO) [17] are used to solve detection and recognition issues.

In this paper, we are using a YOLO V5 Model to detect the objects with the help of camera. In addition, requires a speech generation tool for generating detected images text to speech. This can be achieved with gTTS (google Text To Speech) and pyttsx3 (Text to speech conversion library in python). By overall, the contributions of the proposed work as follows:

- Proposed Third Eye for visually impaired people.
- Used YOLOv5 for image detection on a custom dataset including MS COCO 2017 dataset
- Used pyttsx3 and gTTS for image detection text to speech generation

Henceforth, the paper is coordinated as follows: Section 2 discusses the related work. The Proposed model Third Eye is explained & implemented with the dataset, existing algorithms in Section 3. Section 4 describes the Experimental results of proposed Third Eye using YOLOv5. The result of YOLOv5 processed with text to speech generation algorithms were explained in Section 5. Finally, Section 6 concludes the paper.

## 2. Literature Survey

Many works have been done on making life better for the visually impaired. There is various equipment for the visually impaired, such as sensor-powered walking sticks, speaking calculators, etc. Most of the works are developed with the help of Camera to capture a image or video, pre-process that, and detect the objects with the help of machine or deep learning algorithms. In addition to detection, detected images should be know to the visually impaired people, for this text to speech techniques are used in some literature.

Rajwani et al. [26] presented a system where the input is taken through the camera as a image, is pre-processed using OpenCV, and then the classification and identification is done in Cloud Vision API. Elmannai, Wafa M and Khaled M. Elleithy [9] proposed a system for object detection, where two camera sensors are used to capture the images and analyzed using computer vision methods. Bashiri S et al. [5] proposed a system where the input is taken through a Google Glass Device, then classification and identification are done using Support Vector Machine (SVM) Algorithm. Gianani et al. [10] came up with a system where the image is captured through a camera and pre-processed using OpenCV. Nishajith et al. [19] suggested a framework that uses Raspberry Pi with pre-trained CNN network. The 'ssd_mobilenet_v1_coco_11_06_2017' pre-trained object detection model is used to classify the objects and text to speech conversion is done using *eS peak*. Patel et al. [22] presented a technology where the image is captured through a webcam and used SVM to identify the objects. Tosun, Selman and Enis Karaarslan [32] proposed a system where the image is captured using the android platform and Tiny YOLO is used for object detection which gives the audio output.

In 2019, Wong et al. [33] proposed a real-time CNN-based object identification system for visually impaired people. The object group was filmed in real-time with a webcam, and the picture function was turned off. Then, to detect the sight of visually handicapped people, a sound-based detector was devised. Nasreen, Jawaid et al. [18] presented a system for guiding visually impaired people through the process of item detection. The developed method imports a picture from the back camera into a website and sends it to the server, where the YOLO model is utilised to recognise the objects on the server side. Arjun et al. [20] presented a technology that is wearable with smart glasses and shoes. Both smart shoes and glasses detect the obstacle and pass an audio output to the user. Rahman, Ferdousi et al. [24] developed a visually impaired object detection model with the help of YOLO algorithm and MTCNN are used for

object identification and facial recognition, respectively. Samkit et al. [28] compared different algorithms to detect multiple objects and they found that Haar Cascade is the fastest and CNN gives more accuracy.

Afif, Mouna et al. [1] in 2020, introduced YOLO v3, on a custom dataset that has 16 indoor object classes. They attained 73.19% mAP, they focused on indoor navigation. Later a framework on deep CNN "RetinaNet" for detecting indoor objects was proposed [2], which showed better results than their earlier work. Bhole, Swapnil and Aniket Dhok [6] proposed a transfer learning on Single-Shot Detection (SSD) mechanism for object detection, and implemented it for human as well as currency detection. They achieved 90.2% accuracy on currency detection. Yohannes, Ervin et al. [35] introduced a method to assist the visually impaired around an outdoor environment. They designed a model using DarkNet-53 as a backbone, input is taken from a ZED stereo camera, and the model is trained on PASCAL VOC and MS COCO datasets. Rashika et al. [13] mentioned a method using Mobile-Net SSD, and the images are taken using Jetson Nano, and PiV2 camera, and trained on PASCAL VOC dataset.

Atikur Rahman and Sheikh Sadi [25] proposed an IoT-enabled Automated Object Recognition using SSD Model, SIFT, and MS COCO dataset in 2021. Balachandar, Santhosh et al. [4] developed a technique in which a multi-view object tracking (MVOT) system is employed to address several cameras monitoring and capturing videos. A powerful and precise framework is created with help of the information provided by the videos. Mansi Mabendru and Sanjay Kumar Dubey [17] created a system employing two separate algorithms, YOLO and YOLO v3, and tested accuracy and performance. The SSD Mobile Net model is utilised in the YOLO Tensor flow, The Darknet model is used in YOLO v3. The python library gTTS is used to transform sentences into audio for the audio Feedback. A wearable device with a virtual assistant system, consisting of a total of five components, was proposed by Kanchan Patil et al. [23]. By using hardware buttons and user-provided voice-over commands, these elements can be browsed. Natural language processing is needed for the task of creating voice-based image captions, which was presented by Mohana Priya et al. [3]. The greatest choice for this project is a CNN and LSTM combo; the main objective of the proposed study work is to create the ideal caption for an image. Sandeep Pandasupuleti et al. [21] proposed Voice Translation and Image Recognition using VCC, LSTM, and Flickr_8k dataset. The Table 1 describes about the methods and pros & cons discussed in literature. Still, there is a scope to provide better methods to

Table 1: Trends & Technologies discussed in literature

| **Paper Title** | **Methods** | **Pros & Cons** |
|---|---|---|
| Proposed System on Object Detection for Visually Impaired People. [26] | Android Camera, OpenCV, Google Cloud Vision API, Compare it with Microsoft COCO Dataset and give output. | Since the output is through Android application, it should have enough battery. |
| A Highly Accurate and Reliable Data Fusion Framework for Guiding the Visually Impaired. [9] | Two camera Sensors, Computer Vision Methods, Oriented FAST and Rotated BRIEF (ORB) and KNN Algorithm. | Accuracy of 96%, Used a motherboard connected with various sensors like gyro, compass, GPS, music, FEZ Spider board. |
| Object Detection to Assist Visually Impaired People: A Deep Neural Network Adventure. [5] | Marshfield Clinic Dataset, Google Glass Device, CNN Model, Support Vector Machine Algorithm | Limited number of objects (ex: doors, stairs, signs etc.,) Accuracy over 98% |
| JUVO - An Aid for the Visually Impaired [10] | Camera,Image Capturing and Preprocessing, Object detection Using OpenCV, SSD Framework, MobileNet Architecture | Few objects in Dataset. Indoor Environment, Accuracy of 99.61% |
| Multisensor – based Object Detection in Indoor Environment for Visually Impaired People [22] | USB Webcam, Preprocessing, Statistical Analysis, SVM Classifier. | It can be used for outdoor environment but it is tested for indoor environment only. |

Continued on next page ...

**Table1 Trends & Technologies discussed in literature ... Contd.**

| Paper Title | Methods | Pros & Cons |
|---|---|---|
| Real-Time Object Detection Application for Visually Impaired People: Third Eye. [32] | Camera, OpenCV Processing, Tiny YOLO TensorFlow, Audio Output, COCO Dataset | Only 20 classes in the dataset, Manual selection. |
| Convolutional Neural Network for Object Detection System for Blind People. [33] | Cnn, Used edge box algorithm, Caffnet model, softmax Cifar10 dataset has been used | The object detection models faced difficulty in classifying the object from a picture of ultimate scale |
| Object Detection and Narrator for Visually Impaired People. [18] | Used YOLO. It narrates to the user. It was trained on Imagenet dataset | Results showed that the accuracy is varying depending on phone camera quality and the light effects. iPhone and Samsung have better results than others. |
| Smart Assistive Navigation Devices for Visually Impaired People. [20] | Open CV, Image processing, Used Smart glass and shoes | Both the devices have been developed by using simple, cheap sensors. Their motive is to make both the devices as a part of the user's regular and frequently used objects. |
| An Assistive Model for Visually Impaired People using YOLO and MTCNN [24] | Open CV, YOLO algorithm, Deep learning | The accuracy rate for the object detection procedure was 63–80%, and processing speed was 6-7 FPS. |
| Research on Small Target Detection in Driving Scenarios Based on Improved YOLO Network. [34] | YOLO v3, 2080 Ti machine, Dataset used is Apollo Scape (Baidu's autopilot dataset). | Improvised YOLO v3 and it showed better results compared to YOLO v3. Accuracy is 84.76%. |
| Object Recognition and Classification System for Visually Impaired. [13] | MobileNetSSD (SSD - Single Shot-Detector), PASCAL VOC 2007. | Got pretty good accuracy, but the dataset is small, not sufficient. Only for embedded systems. |
| An Evaluation of RetinaNet on Indoor Object Detection for Blind and Visually Impaired Persons Assistance Navigation. [2] | RetinaNet (ResNet, DenseNet, VGGNet based), Self prepared Dataset (Contains 8000 images). | Attained 84.61% mAP. Focused on only indoor navigation. the number of objects it can detect is very small. Got good results with proposed algorithm. |
| Indoor object detection and recognition for an ICT mobility assistance of visually impaired people. [1] | YOLO v3, DarkNet-53. Dataset contains 8000 images and contains 16 indoor object classes. | Attained 73.19% mAP, and it's only focused on indoor navigation. Used pretrained model and trained on the new dataset. |
| Robot Eye: Automatic Object Detection and Recognition Using Deep Attention Network to Assist Blind People. [35] | Self-designed model (DarkNet-53 based), ZED Stereo camera, PASCAL VOC, MS COCO datasets. | Accuracy is 81%, better than YOLO v3. Used PASCAL VOC for classes, and mixed MS COCO. No-of classes are too small. |
| Deep Learning based Object Detection and Recognition Framework for the Visually-Impaired. [6] | PASCAL VOC 2007 dataset, SSD, Inception v3 model. | Added currency detection to the dataset and achieved 90.2% acc. But the dataset contains only 20 classes. |
| IoT Enabled Automated Object Recognition for the Visually Impaired. [25] | laser sensors, Single Shot Detector (SSD) model, SIFT, MS COCO dataset | YOLO accuracy is 95.99% and SSD 88.89%. YOLO seems to be better compare to SSD. |

Continued on next page ...

**Table1 Trends & Technologies discussed in literature ... Contd.**

| Paper Title | Methods | Pros & Cons |
|---|---|---|
| Deep Learning Technique Based Visually Impaired People Using YOLO v3 Framework Mechanism. [4] | YOLO v3, Cameras, MVOT, COCO dataset | They have used (videocon camera) its intra camera grahic. which does not highlights the features properly and exactly tally the model. |
| Real Time Object Detection with Audio Feedback using Yolo vs. Yolo_v3 [17] | Tensor flow, SSD, YOLO v3, gTTS, Deep Learning | YOLO accuracy is 78.99 and YOLO v3 92.89% (seems to be better compare to YOLO). |
| Guidance System for Visually Impaired People. [23] | gTTS, YOLO v3, Pyttsx, AIML, Vice over chatbot | In a noisy environment, the command may be misinterpreted by the chatbot as coming from a nearby person. |
| Building A Voice Based Image Caption Generator with Deep Learning. [3] | NLP ,CNN, LSTM (Long short term memory), RNN (recurrent neural network) flicker dataset, Accuracy 90% | The dataset is small. For better accuracy could be used big dataset, According to current trends, it's not sufficient. |
| Image Recognition and Voice Translation for Visually Impaired. [21] | Flickr_8k dataset, VGG, LSTM | Dataset is very small, the implementation can be enhanced by giving a greater number of images and text datasets with shorter captions for training. |

## 3. The Proposed Work: Third Eye

In this proposed work, images can be captured using camera, which is placed on top of the visually impaired person head. The captured images are passed through the YOLOv5 model. YOLOv5 Model detects the images. These detected images are passed to speech generation module (pyttsx3 or gTTS), which will generate the speech of the objects present in that image (in front of visually impaired person). Explanation of the proposed model is shown in Figure 1.

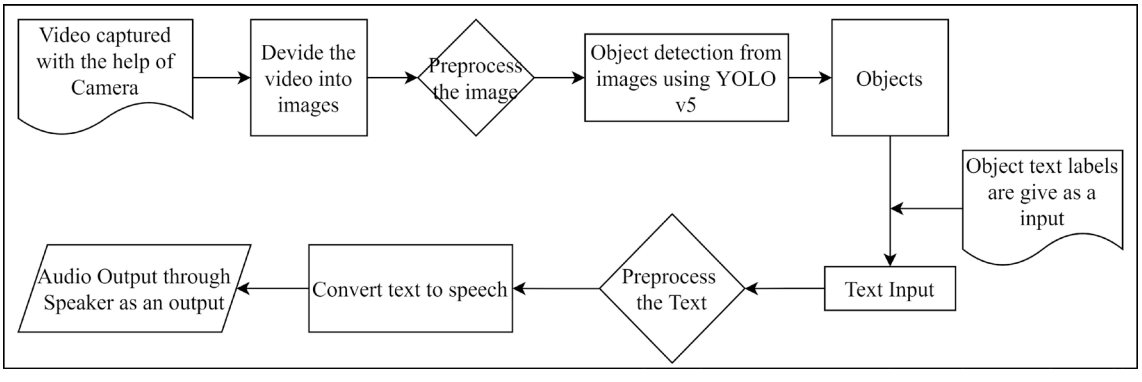


Fig. 1. Schematic diagram of proposed system.

### 3.1. Dataset

Training, Validation, and Testing of proposed model using YOLOv5 are done on a custom prepared dataset combined with MS COCO 2017 Dataset [15]. MS COCO 2017 dataset contains 80 different object classes such as person, dog, chair, potted plant, etc. In addition, we added 15 more different object classes such as switchboard, pillow, locker, keys, open door, closed-door, window, direction board, postbox, pole, shop, manhole, tree, upstairs, downstairs. Which are not mentioned in MS COCO 2017 Dataset (95 classes overall). These objects are relevant to Indian atmosphere. For each object class, we added 30 - 50 images, all together we added 500 images to dataset. By overall 5000 images are considered for doing image detection.

### 3.2. Annotation tool

Used makesense.ai [30] a data annotation tool to annotate new dataset, which contains 15 objects, which are mentioned in Section 3.1. Makesense provides a lot more flexibility in adding labels to the images. It is also possible to download the annotated images in YOLO format. So, this is the reason why we choose makesense.ai as our annotation tool.

### 3.3. Methods used in Proposed Model

The project uses YOLO algorithm that provides real-time object detection. For speech generation from seeing objects, used pyttsx3 or gTTS modules.

#### 3.3.1. YOLO v5 for Object Detection

The "You Only Look Once" (YOLO) object identification approach concentrates on finding things in photographs and grouping them into a grid structure. It is the responsibility of each grid cell to find objects inside its borders. At the moment, YOLO v5 is one of the best object detection models available. The beautiful thing about this Deep Neural Network is that retraining it on our custom dataset is quite simple [31]. About 90% less space is used in the YOLO v5 version compared to the YOLO v4 version. With accuracy on par with the YOLO v4 test, YOLO v5 is claimed to be much faster and lighter than YOLO v4. As a result, we chose YOLO v5.

#### 3.3.2. Text to speech synthesizer

The goal of the text processing component is to process the provided input text and produce a suitable phonemic unit sequence. The incoming text is first processed, normalized, and transcribed into a phonetic or other linguistic representation in a text-to-speech system. Low-level processing difficulties like sentence segmentation and word segmentation are dealt with following text processing components.

- Document Structure detection: The document structure can be detected by diagnosing punctuation marks and paragraph formatting.
- Text Normalization: The text normalization controls abbreviations and acronyms. The goal of normalization is to make the text correspond, for example, Dr could be represented as the doctor. Valid normalization constructs a fair result.
- Linguistic Analysis: Linguistic analysis contains a morphological analysis for syntactic analysis and accurate word pronunciation to promote accenting and phrasing to manage obscurities in written text.

A speech synthesiser is used by the text-to-speech system (TTS) to convert text into voice. It creates a human voice in an artificial way. A computer system used for this is called a voice synthesiser. A text-to-speech system's two primary components are text processing and speech synthesis. Figure 2 from the article [16] depicts the process of text-to-speech synthesis.

Some of the Text to speech conversion libraries using in proposed model are:

- Google Text to Speech (gTTS): gTTS is a programme that turns text into audio files that may be saved as mp3 files. The gTTS API supports English, Hindi, Tamil, French, German, and a variety of additional languages. It

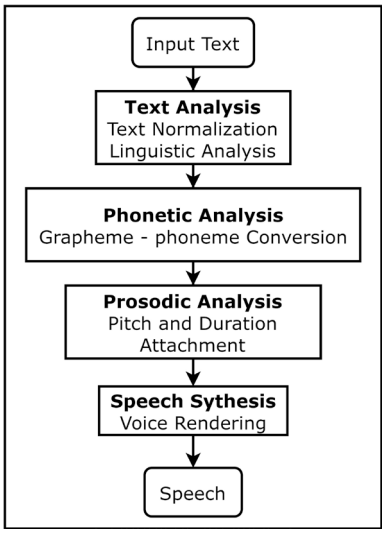


Fig. 2. Text to speech synthesis - block diagram.

includes a speech-specific sentence tokenizer that enables for endless amounts of text to be read while keeping accurate intonation, abbreviations, decimals, and more, as well as customisable text pre-processors that can improve pronunciation, among other things.

- Text to speech conversion library in python (pyttsx3): Pyttsx3 is a Python text-to-speech library. Unlike other libraries, it works both offline and online, and is compatible with both Python 2 and 3 versions. It works without any delay. There are some customization's available. we can change the voice of the engine. We can also change the speed of the voice engine.

## 4. Experimental Results

We carried out training, validation, and testing on the google colab platform. Weights & Biases [7] are used to track the training and validation process for visualization. While Training and Validation, considered the following losses for better understanding of the proposed system advantages.

*Box loss:* The box loss is used to check whether the predicted bounding box covers the object or not. In addition, How well the algorithm can find center of object will be identified with the help of box loss.

*Class loss:* Class loss will be used to determine how effectively the algorithm can predict the correct class for the given item.

*Object loss:* The probability that an object exists in a suggested region of interest is measured by objectness.

### *4.1. Training*

Tesla K80 with 12 GB RAM, powered by google colab is used for training the YOLO v5 model, with the help of PyTorch and PyTorch-Cuda libraries which are coded in python. The model is trained on the dataset mentioned for 50 epochs, With a batch size of 8. Here are the class loss (shown in Figure 3), Box loss (shown in Figure 4), Object loss (shown in Figure 5) results for the training set.

### *4.2. Validation*

Each training epoch is followed by 50 epochs of validation with a batch size of 16 for each training epoch. Here are the results for the validation set for class loss (Figure 6), box loss (Figure 7), and object loss (Figure 8).

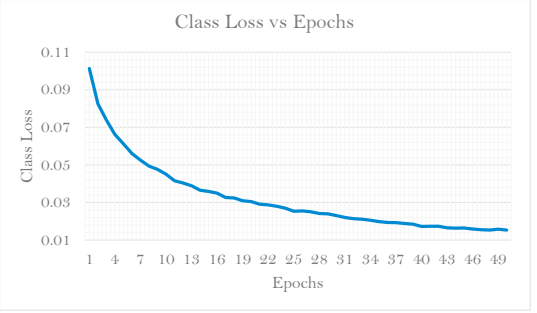


Fig. 3. Training: Class loss vs number of epochs.

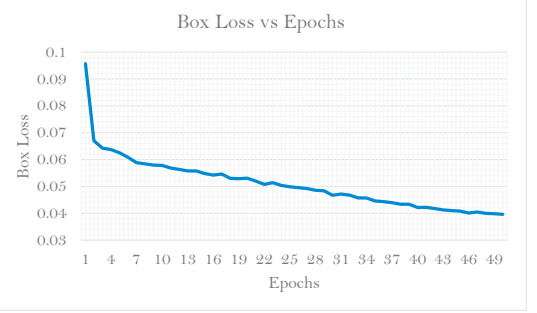


Fig. 4. Training: Box loss vs number of epochs.

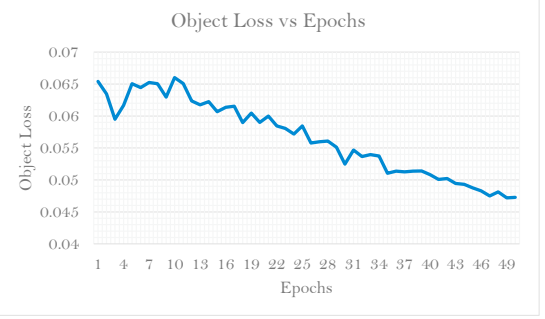


Fig. 5. Training: Object loss vs number of epochs.

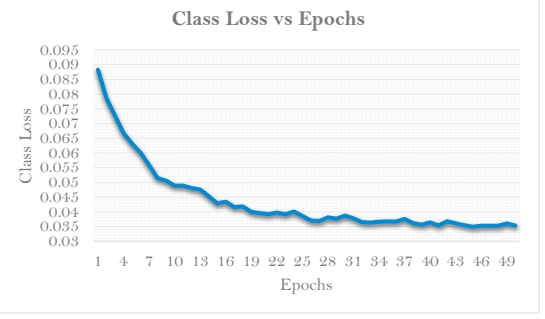


Fig. 6. Validation: Class loss vs number of epochs.

### 4.3. Evaluation metrics

Model is evaluated based on Precision, Recall, mAP (mean Average Precision).

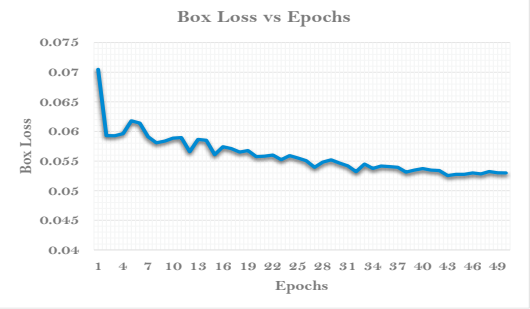


Fig. 7. Validation: Box loss vs number of epochs.

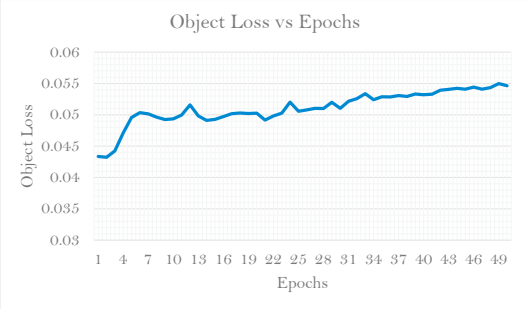


Fig. 8. Validation: Object loss vs number of epochs.

- **Precision:** It is used as a performance measurement for machine learning model. It can be the accuracy of a model's positive prediction. Precision is calculated as the ratio of true positives to all positive predictions. The precision achieved by our model is shown in Figure 9.
- **Recall:** A recall measures how many right positive predictions are made out of all possible positive predictions. Positive predictions that were missed are indicated by the recall. The recall can be calculated and presented in Figure 10.
- **mean Average Precision (mAP):** The mean Average Precision (mAP) score is derived by averaging the AP across all classes and/or the overall Intersection over Union thresholds, depending on the various detection issues that are present.. The mAP of our model is shown in Figure 11 & Figure 12.

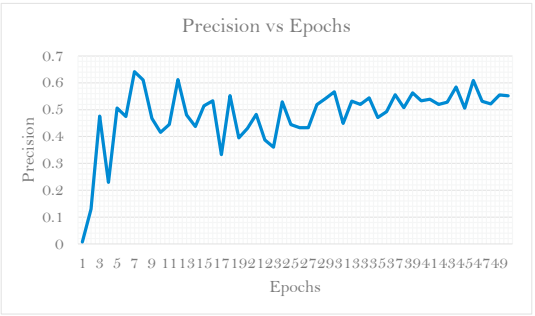


Fig. 9. Evaluation metric: Precision vs number of epochs.

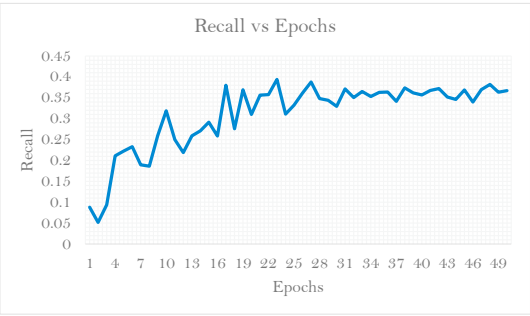


Fig. 10. Evaluation metric: Recall vs number of epochs.

## 5. Output tensor to speech conversion

The output of YOLOv5 is a tensor of objects. The tensor's objects each have six values. I.e., x, y, w, h, label, and confidence. In this case, the detected box's centre is (x, y), its width and height are (w, h), confidence indicates how likely it is that an object is present in the box and how accurate the bounding box is, and label designates the object that was discovered.

To convert the output tensor to speech, divided it into two parts, one is detected objects to text, and the other is text to speech.

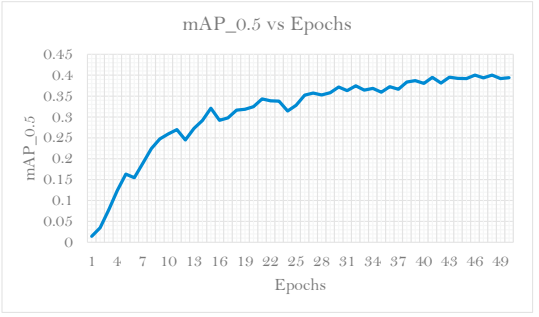


Fig. 11. Evaluation metric: mAP_0.5 vs number of epochs.

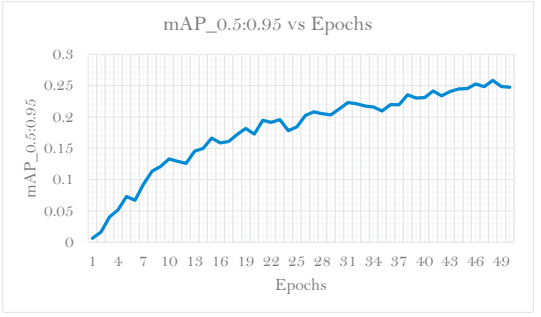


Fig. 12. Evaluation metric: mAP_0.5:0.95 vs number of epochs.

### *5.1. Output tensor to text*

A function is defined to generate text from output tensor, for further speech generation. The function takes the following parameters:

- results - output tensor of YOLOv5.
- H - Height of the window (Image).
- W - Width of the window (Image).
- names - The list of labels, with which the model is trained.

The function iterates over the results. In each iteration, it creates a text describing the position and object and adds to a list of text. The window (Image) is divided into nine parts (three parts - horizontally, three parts vertically, overall it makes nine). Each bounding box contains a center point, with which we find at which place the object lies in the view. Finally, all the text in the list is joined with a comma-separated delimiter, which is then returned.

To show the advantages of proposed approach compared with existing approaches by concentrating on parameters such as model Considered, Data sets used for implementing the approaches, audio feedback methods were used or not is shown in Table 2

## 6. Conclusion

The proposed Third Eye is used YOLOv5 for object detection and pyttsx3 and gTTS for speech generation. The proposed method is able to detect the images and generate speech accurately to help the visually impaired people, who are staying alone in their homes. Proposed model YOLOv5 is able to detect 95 different objects, with high confidence. From the two python libraries for speech generation, we observed that pyttsx3 doesn't require any internet connection, whereas, on the other side, gTTS need constant internet connectivity. To create a voice file, gTTS sends text to Google's servers, which are subsequently returned. The speech generated by pyttsx3 is spoken comparatively faster than gTTS. Hence, we found pyttsx3 more helpful than gTTS, considering the time taken to produce audio, the delay in frames, the libraries required, and the network connectivity. Since, we want to develop a device that should not be affected by any external factors like bad network, etc. So, we used pyttsx3 as our main library.

In Future, we planing to recognize the persons who are visiting homes. This will give better safety for visually impaired people with absence of caretakers. In addition, planning to send the visitor images to the care taker.

## References

[1] Afif, M., Ayachi, R., Pissaloux, E., Said, Y., Atri, M., 2020a. Indoor objects detection and recognition for an ict mobility assistance of visually impaired people. Multimedia Tools and Applications 79, 31645–31662.

Table 2. Comparison of proposed approach with existing approaches

| Models | Dataset used | Model used | Audio feedback implemented | Audio feedback methods | Drawbacks |
|---|---|---|---|---|---|
| [1] | Can detect only 16 indoor objects with 8000 images | YOLO v3 | NO | - | images limited to indoor |
| [3] | Flicker dataset (5000 Images) | Combination of CNN and LSTM | YES | LSTM | Architecture not mentioned. Its a small model, cannot extract more features |
| [23] | Not mentioned | Trained a CV model | YES | A voice-over assistant module | In a noisy environment, the command may be misinterpreted by the chatbot as coming from a nearby person. |
| [24] | COCO and ImageNet, Custom made dataset | YOLO v2, MTCNN | YES | Not mentioned | Audio is not provided as a output |
| [32] | COCO dataset (20 classes) | Tiny YOLO | NO | Not mentioned | Only 20 classes in the dataset. The model strength is compromised for making it useful for mobile. Audio is not provided as a output |
| Proposed Model | Custom prepared dataset combined with MS COCO 2017 Dataset | Used YOLO v5 | YES | Compared gTTS, pyttsx3 | Advantages: faster and light weight model, completely implemented with audio feedback. |


[2] Afif, M., Ayachi, R., Said, Y., Pissaloux, E., Atri, M., 2020b. An evaluation of retinanet on indoor object detection for blind and visually impaired persons assistance navigation. Neural Processing Letters , 1–15.

[3] Anu, M., Divya, S., et al., 2021. Building a voice based image caption generator with deep learning, in: 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE. pp. 943–948.

[4] Balachandar, A., Santhosh, E., Suriyakrishnan, A., Vigensh, N., Usharani, S., Bala, P.M., 2021. Deep learning technique based visually impaired people using yolo v3 framework mechanism, in: 2021 3rd International Conference on Signal Processing and Communication (ICPSC), IEEE. pp. 134–138.

[5] Bashiri, F.S., LaRose, E., Badger, J.C., D'Souza, R.M., Yu, Z., Peissig, P., 2018. Object detection to assist visually impaired people: A deep neural network adventure, in: International Symposium on Visual Computing, Springer. pp. 500–510.

[6] Bhole, S., Dhok, A., 2020. Deep learning based object detection and recognition framework for the visually-impaired, in: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), IEEE. pp. 725–728.

[7] Biewald, L., 2020. Experiment tracking with weights and biases. URL: https://www.wandb.com/. software available from wandb.com.

[8] Center, M.E., 2016. Importance of Eye Care, available at, https://www.medicaleyecenter.com/2016/06/20/importance-eye-care/. URL: https://www.medicaleyecenter.com/2016/06/20/importance-eye-care/. [Online; accessed 6-November-2021].

[9] Elmannai, W.M., Elleithy, K.M., 2018. A highly accurate and reliable data fusion framework for guiding the visually impaired. IEEE Access 6, 33029–33054.

[10] Gianani, S., Mehta, A., Motwani, T., Shende, R., 2018. Juvo-an aid for the visually impaired, in: 2018 International Conference on Smart City and Emerging Technology (ICSCET), IEEE. pp. 1–4.

[11] Guravaiah, K., Rithika, G., Raju, S.S., 2022. Homeid: Home visitors recognition using internet of things and deep learning algorithms, in: 2022 International Conference on Innovative Trends in Information Technology (ICITIIT), IEEE. pp. 1–4.

[12] Guravaiah, K., Velusamy, R.L., 2019. Prototype of home monitoring device using internet of things and river formation dynamics-based

multi-hop routing protocol (rfdhm). IEEE Transactions on Consumer Electronics 65, 329–338.
[13] Joshi, R., Tripathi, M., Kumar, A., Gaur, M.S., 2020. Object recognition and classification system for visually impaired, in: 2020 International Conference on Communication and Signal Processing (ICCSP), IEEE. pp. 1568–1572.
[14] Kumar, R., Singh, A., Datta, G., Kumar, A., Garg, H., 2021. Brain tumor detection system using improved convolutional neural network, in: 2021 Sixth International Conference on Image Information Processing (ICIIP), IEEE. pp. 126–130.
[15] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: European conference on computer vision, Springer. pp. 740–755.
[16] Mache, S.R., Baheti, M.R., Mahender, C.N., 2015. Review on text-to-speech synthesizer. International Journal of Advanced Research in Computer and Communication Engineering 4, 54–59.
[17] Mahendru, M., Dubey, S.K., 2021. Real time object detection with audio feedback using yolo vs. yolo_v3, in: 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE. pp. 734–740.
[18] Nasreen, J., Arif, W., Shaikh, A.A., Muhammad, Y., Abdullah, M., 2019. Object detection and narrator for visually impaired people, in: 2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS), IEEE. pp. 1–4.
[19] Nishajith, A., Nivedha, J., Nair, S.S., Shaffi, J.M., 2018. Smart cap-wearable visual guidance system for blind, in: 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE. pp. 275–278.
[20] Pardasani, A., Indi, P.N., Banerjee, S., Kamal, A., Garg, V., 2019. Smart assistive navigation devices for visually impaired people, in: 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), IEEE. pp. 725–729.
[21] Pasupuleti, S., Dadi, L., Gadi, M., Krishnaveni, R., 2021. Image recognition and voice translation for visually impaired. International Journal of Research in Engineering, Science and Management 4, 18–23.
[22] Patel, C.T., Mistry, V.J., Desai, L.S., Meghrajani, Y.K., 2018. Multisensor-based object detection in indoor environment for visually impaired people, in: 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE. pp. 1–4.
[23] Patil, K., Kharat, A., Chaudhary, P., Bidgar, S., Gavhane, R., 2021. Guidance system for visually impaired people, in: 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), IEEE. pp. 988–993.
[24] Rahman, F., Ritun, I.J., Farhin, N., Uddin, J., 2019. An assistive model for visually impaired people using yolo and mtcnn, in: Proceedings of the 3rd International Conference on Cryptography, Security and Privacy, pp. 225–230.
[25] Rahman, M.A., Sadi, M.S., 2021. Iot enabled automated object recognition for the visually impaired. Computer Methods and Programs in Biomedicine Update , 100015.
[26] Rajwani, R., Purswani, D., Kalinani, P., Ramchandani, D., Dokare, I., 2018. Proposed system on object detection for visually impaired people. International Journal of Information Technology (IJIT) 4, 1–6.
[27] Rastogi, P., Singh, V., Yadav, M., 2018. Deep learning and big datatechnologies in medical image analysis, in: 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), IEEE. pp. 60–63.
[28] Shah, S., Bandariya, J., Jain, G., Ghevariya, M., Dastoor, S., 2019. Cnn based auto-assistance system as a boon for directing visually impaired person, in: 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), IEEE. pp. 235–240.
[29] Singh, V., Asari, V.K., Rajasekaran, R., 2022. A deep neural network for early detection and prediction of chronic kidney disease. Diagnostics 12, 116.
[30] Skalski, P., 2019. Make Sense available at, https://www.makesense.ai/. URL: https://www.makesense.ai/. [Online; accessed 6-November-2021].
[31] Tan, M., Pang, R., Le, Q.V., 2020. Efficientdet: Scalable and efficient object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10781–10790.
[32] Tosun, S., Karaarslan, E., 2018. Real-time object detection application for visually impaired people: Third eye, in: 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Ieee. pp. 1–6.
[33] Wong, Y.C., Lai, J., Ranjit, S., Syafeeza, A., Hamid, N., 2019. Convolutional neural network for object detection system for blind people. Journal of Telecommunication, Electronic and Computer Engineering (JTEC) 11, 1–6.
[34] Xu, Q., Lin, R., Yue, H., Huang, H., Yang, Y., Yao, Z., 2020. Research on small target detection in driving scenarios based on improved yolo network. IEEE Access 8, 27574–27583.
[35] Yohannes, E., Lin, P., Lin, C.Y., Shih, T.K., 2020. Robot eye: Automatic object detection and recognition using deep attention network to assist blind people, in: 2020 International Conference on Pervasive Artificial Intelligence (ICPAI), IEEE. pp. 152–157.