Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2023) 000–000

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

Visually impaired people face a lot of difficulties in doing their daily activities. 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.Visually impaired need the help of either the third person or a stick. These methods are not always fruitful. Detecting and recognizing the objects and generating speech about the objects helps visually impaired in a great way in understanding their surroundings. To assist the visually impaired 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 Third Eye, giving better accuracy in detection and speech generation to help the visually impaired people.

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

1. Introduction

Visually Impaired face a lot of difficulties in their daily lives. 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.

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

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

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. The deep learning systems achieve high accuracy rates at a lower cost. Many Convolutional Neural Network (CNN) methods such as Single Shot Detector (SSD) [14] and You Only Look Once (YOLO) [18] are used to solve detection and recognition issues. There are other architectures such as Faster R-CNN and Mask R-CNN [36]. 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.

Rajwani, Roshan et al. [27] presented a system where the input is taken through an android camera, then the captured image is preprocessed using OpenCV, then the classification and identification is done in Cloud Vision API. Elmannai, Wafa M and Khaled M. Elleithy [10] proposed a system for object detection, where two camera sensors are used, which are then analyzed using computer vision methods. The ORB and KNN are used for object detection. Ye, Cang and Xiangfei Qian [35] in 2018, a 3-D Object Recognition for Visually Impaired people is proposed. The cane used by blind people is attached with a CV enhanced 3D Camera, it captures a 3D point cloud which is segmented into planar segments, which are then classified using Gaussian Model Mixture and clustered into the target objects. Bashiri, Fereshteh S et al. [6] proposed a system where the input is taken through a Google Glass Device, then classification and identification are done using Support Vector Machine Algorithm. Gianani, Sejal et al. [12] came up with a system where the image is captured through a camera device for the input and preprocessed using OpenCV. They used the SSD framework in conjunction with the MobileNet architecture. Nishajith, A et al. [20] suggested a framework that uses Raspberry Pi which has a Pre-trained CNN network. The image is captured through Noir Camera and preprocessing is done through OpenCV and they used Pre-trained object detection model 'ssd_mobilenet_v1_coco_11_06_2017' to classify the objects and text to speech conversion is done using eSpeak. Patel, Charmi T et al. [23] presented a technology where the image is captured through a USB webcam and preprocessing is done and it classifies and identifies the objects using the SVM Algorithm. Tosun, Selman and Enis Karaarslan [32] proposed a system where the image is captured using the android platform and preprocessing is done using OpenCV and Tiny YOLO is used for object detection which gives the audio output.

Wong, Yan Chiew et al. [33] In 2019, a real-time CNN-based object identification system for visually impaired people was proposed. 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. [19] 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. Pardasani, Arjun et al. [21] presented a technology that is wearable like smart glasses and shoes. Both smart shoes and glasses detect the obstacle and pass an audio output to the user. Rahman, Ferdousi et al. [25] developed a visually impaired object detection model based on the YOLO algorithm. For the building model, MTCNN is used. The YOLO Algorithm and MTCNN Networking are used for object identification and facial recognition, respectively. Shah, Samkit et al. [28] compared different detection algorithms to detect multiple objects and they found that Haar Cascade is the fastest and CNN gives more accuracy. Jhinkwan, Piyush et al. [13] proposed a system that uses a convolutional network combined with fully connected layers. Chen, Xiaobai et al. [9] created an automatic DCNN quantization approach to decrease the data range to 4 or 5 bits. Sun, Minghui et al. [30] presented a data collection system based on Google Tango, which has an infrared (IR) sensor built in.

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. Afif, Mouna et al. [2] later proposed a framework on deep CNN "RetinaNet" for detecting indoor objects, which showed better results than their earlier work. Fang, Wei et al. [11] introduced a method using the Tinier-YOLO model, which is 4 times smaller than Tiny-YOLO v3. trained on PASCAL VOC and COCO datasets. It's faster than other lightweight models. Li, Yongjun et al. [15] proposed another version of YOLO, that is YOLO-ACN, which showed better results. They mainly focused on small objects detection. Bhole, Swapnil and Aniket Dhok [7] 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. [36] 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. Joshi, Rashika et al. [14] mentioned a method using Mobile Net SSD, and the images are taken using Jetson Nano, and PiV2 camera, and trained on PASCAL VOC dataset. Achieved pretty good results with the proposed model.

Atikur Rahman and Sheikh Sadi [26] proposed an IoT-enabled Automated Object Recognition using SSD Model, SIFT, and MS COCO dataset in 2021. Balachandar, Santhosh et al. [5] developed a technique in which a multi-view object tracking (MVOT) system is employed to address several cameras monitoring and capturing videos in this proposed system. And, by merging the information from the videos, a powerful and precise framework is created. Using the YOLO v3 algorithm, each segmented group of objects in one view is mapped to the equivalent group in another view. Blob gathers, which allow data to be transferred across cameras, corresponded to these agreeing sets. After being taken by the camera, these visuals are converted into vocal output. Mansi Mabendru and Sanjay Kumar Dubey [18] 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. Kanchan Patil et al. [24] proposed a wearable device with a virtual assistant system for visually challenged people, with a total of five components integrated into one system. These components can be navigated via hardware buttons and voice-over commands provided by the user. Mohana Priya et al. [4] presented a voice-based image caption generation, which is a task that requires the use of natural language processing.The best option in this project is a combination of CNN and LSTM; the major goal of this proposed study work is to produce the perfect caption for an image. The description will be transformed into text, and the text will be converted into a voice. You Only Look Once (YOLO) a Real-Time Object Detection is deployed by Annapoorani et al. [3] proposed a model where the image features are identified using image classification techniques, and the Indian money identification module is utilised to identify the denominations. Using the gTTS package, the written description of the identified object will be transmitted to the gTTS API. Sandeep Pandasupuleti et al. [22] proposed Voice Translation and Image Recognition using VCC, LSTM, and Flickr_8k dataset. The following table (Table 1) describes about the methods and pros & cons discussed in literature.

Table 1: Trends & Technologies discussed in literature

Paper Title	**Methods**	**Pros & Cons**
Proposed System on Object Detection for Visually Impaired People. [27]	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. [10]	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.

Continued on next page

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

Paper Title	Methods	Pros & Cons
3-D Object Recognition of a Robotic Navigation Aid for the Visually Impaired [35]	3D Camera(White Cane), Planar Segments, Gaussian Model Mixture	Trained on all indoor objects Accuracy over 90%
Object Detection to Assist Visually Impaired People: A Deep Neural Network Adventure. [6]	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 [12]	Camera,Image Capturing and Preprocessing, Object detection Using OpenCV, SSD Framework, MobileNet Architecture	Few objects in Dataset. Indoor Environment, Accuracy of 99.61%
Smart Cap-Wearable Visual Guidance System For Blind. [20]	Raspberry Pi Noir Camera, OpenCV Processing, COCO Model, eSpeak.	90 classes of objects in Dataset.
Multisensor – based Object Detection in Indoor Environment for Visually Impaired People [23]	USB Webcam, Preprocessing, Statistical Analysis, SVM Classifier.	It can be used for outdoor environment but it is tested for indoor environment only.
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. [19]	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. [21]	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 [25]	Open CV, YOLO algorithm, Deep learning	The object detection process achieved 6-7 FPS processing with an accuracy rate of 63-80.%
CNN based Auto-Assistance System as a Boon for Directing Visually Impaired Person. [28]	Haar cascade, CNN, Deep learning COCO 2017 data Set was used	When processed on CPU, Haar cascade is the fastest algorithm, but CNN gives more accurate results when detecting multiple objects simultaneously for real time applications.
Object Detection Using Convolution Neural Networks [13]	Deep learning, CNN, Back propagation algorithm. For training CIFAR-100 dataset was used	It was trained with dropout and data augmentation to achieve better results.

Continued on next page

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

Paper Title	Methods	Pros & Cons
A 68 mw 2.2 Tops/w low bit-width and multiplierless DCNN object detection processor for visually impaired people. [9]	Deep convolutional network, low-bit, multiplierless	reducing hardware cost by over 68% compared to the 16 bit fix-point model with negligible accuracy loss.
”Watch Your Step”: Precise Obstacle Detection and Navigation for Mobile Users Through Their Mobile Service [30]	Google Tango, built-in infrared (IR) sensor to collect data	The system cannot correctly distinguish complex situations such as obstacles leaning against a wall.
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%.
Tinier-YOLO: A Real-Time Object Detection Method for Constrained Environments. [11]	Tinier-YOLO-v3, PASCAL VOC (2007 + 2012), COCO.	Faster runtime speed compared to other lightweight models. But, is suitable for embedded systems (Low accuracy).
YOLO-ACN: Focusing on Small Target and Occluded Object Detection. [15]	YOLO-ACN, MS COCO, Infrared pedestrian dataset KAIST, NVIDIA Tesla K40.	Doesn't improve performance much with the proposed method, compared to YOLO v3. focused on small objects detection.
Object Recognition and Classification System for Visually Impaired. [14]	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. [36]	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. [7]	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. [26]	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.
Deep Learning Technique Based Visually Impaired People Using YOLO v3 Framework Mechanism. [5]	YOLO v3, Cameras, M VOT, COCO dataset	They have used (videocon camera) its intra camera grahic. which does not highlights the features properly and exactly tally the model.

Continued on next page

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

Paper Title	Methods	Pros & Cons
Real Time Object Detection with Audio Feedback using Yolo vs. Yolo_v3 [18]	Tensor flow, SSD, YOLO, 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. [24]	gTTS, YOLO v3, Pyttsx, AIML, Vice over chatbot	chat-bot cannot recognize the command in noisy environment, chat-bot may get confused between voice of an user and person nearby.
Building A Voice Based Image Caption Generator with Deep Learning. [4]	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.
Blind - Sight: Object Detection with Voice Feedback. [3]	YOLO, COCO Dataset, gTTS	Live object recognition system cannot perform future learning which is a demerit.
Image Recognition and Voice Translation for Visually Impaired. [22]	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 propsed work, The 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. Explanation of the proposed model is shown in Figure 1.

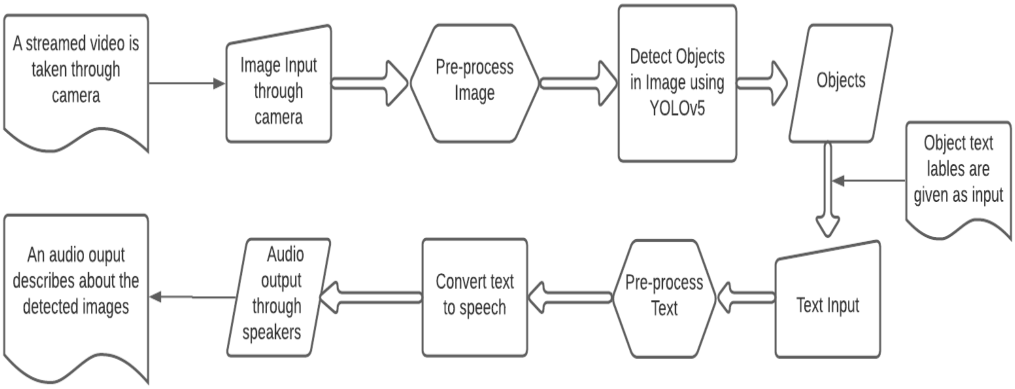
A streamed video is taken through camera
Image Input through camera
Pre-process Image
Detect Objects in Image using YOLOv5
Objects
Object text lables are given as input
Text Input
Pre-process Text
Convert text to speech
Audio output through speakers
An audio ouput describes about the detected images

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 [16]. MS COCO 2017 dataset contains 80 different object classes likely, person, dog, chair, potted plant, etc. In addition, we added 15 more different object classes such as switchboard, pillow, locker, keys, open door, closeddoor, 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 [29] 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 than other tools in adding labels list, most of the other tools automatically order the labels alphabetically. But, makesense follows the order we provide, and 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

The object identification method YOLO, which stands for ”You Only Look Once,” focuses on detecting objects in photos and separates them into a grid structure. Each grid cell is in charge of detecting items within its boundaries. 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].

The YOLO v5 version is roughly 90% smaller than the YOLO v4 version. As a result, YOLO v5 is touted to be much faster and lighter than YOLO v4, with accuracy comparable to the YOLO v4 test. As a result, we chose YOLO v5.

3.3.2. Text to speech synthesizer

The text processing component aims to process the given input text and generate a phonemic unit sequence that is appropriate. 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.

Text to speech system (TTS) transforms text into a voice using a speech synthesizer. It artificially produces a human voice. A speech synthesiser is a computer system that is used for this purpose. Text processing and speech generation are two main elements of a text-to-speech system. The process of Text to speech synthesis is shown in Figure 2 taken from literature [17].

Some of the Text to speech conversion libraries using in propose model

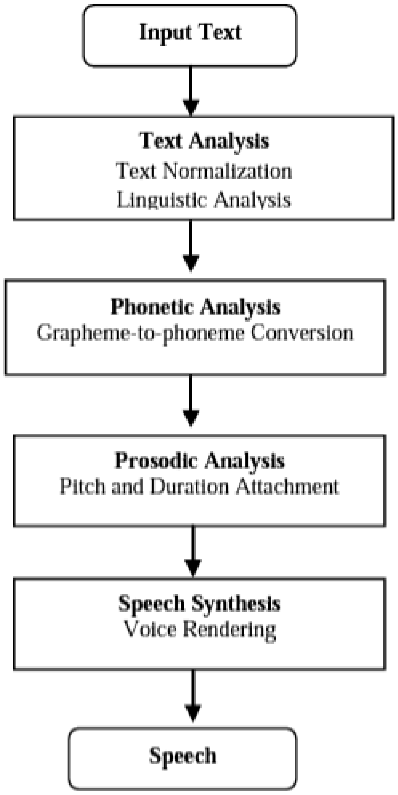

Fig. 2. Text to speech synthesis - block diagram.

- Google Text to Speech (gTTS).
- Text to speech conversion library in python (pyttsx3).

3.3.3. 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 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.

3.3.4. 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 our training, validation, and testing on the google colab platform. Weights & Biases [8] is used to track the training and validation process for visualization. While Training and Validation, considered following losses for better understanding of the proposed system. *Box loss:* The box loss represents how well the algorithm can locate the center of an object and how well the predicted bounding box covers an object.

Class loss: Classification loss gives an idea of how well the algorithm can predict the correct class of a given object.

Object loss: Objectness is essentially a measure of the probability that an object exists in a proposed region of interest.

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 6), Box loss (shown in Figure 7), Object loss (shown in Figure 8) results for the training set.

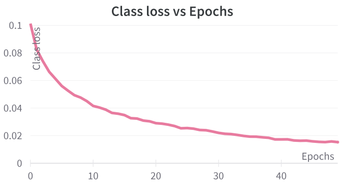

Fig. 3. Training: Class loss vs number of epochs.

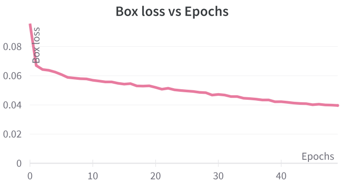

Fig. 4. Training: Box loss vs number of epochs.

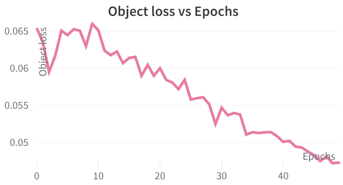

Fig. 5. Training: Object loss vs number of epochs.

4.2. Validation

Validation is done on each training epoch with a batch size of 16, for 50 epochs after each training epoch. Here are the class loss (shown in Figure 9), Box loss (shown in Figure 10), Object loss (shown in Figure 11) results for the validation set.

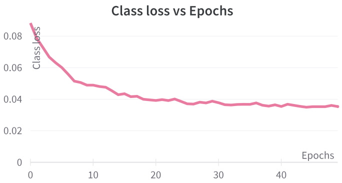

Fig. 6. Validation: Class loss vs number of epochs.

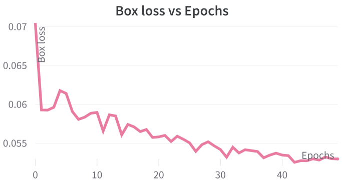

Fig. 7. Validation: Box loss vs number of epochs.

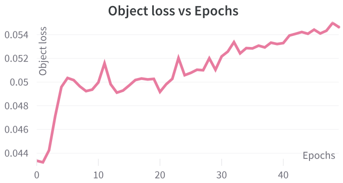

Fig. 8. Validation: Object loss vs number of epochs.

4.3. Evaluation metrics

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

- **Precision:** Precision is one measure of a machine learning model's performance – the accuracy of a model's positive prediction. The number of true positives divided by the total number of positive predictions is known as precision. The precision achieved by our model is shown in Figure 12.
- **Recall:** A recall is a metric that measures how many right positive predictions were made out of all possible positive predictions. Positive predictions that were missed are indicated by the recall. The recall achieved by our model is shown in Figure 13.
- **mean Average Precision (mAP):** Depending on the different detection problems that exist, the mean Average Precision or mAP score is calculated by taking the mean AP over all classes and/or overall IoU thresholds. The mAP of our model is shown in Figure 14 & Figure 15.

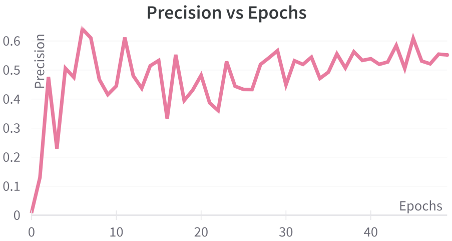

Fig. 9. Evaluation metric: Precision vs number of epochs.

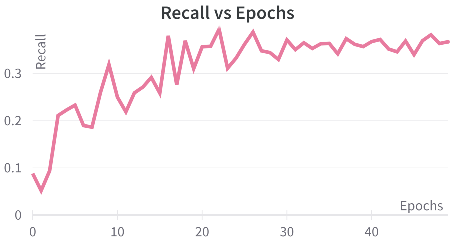

Fig. 10. Evaluation metric: Recall vs number of epochs.

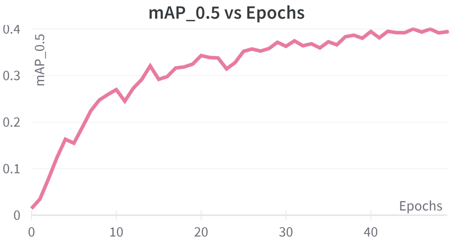

Fig. 11. Evaluation metric: mAP_0.5 vs number of epochs.

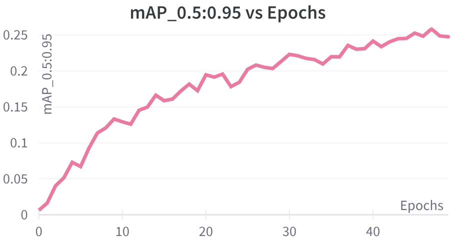

Fig. 12. Evaluation metric: mAP_0.5:0.95 vs number of epochs.

5. Output tensor to speech conversion

The output of YOLOv5 is a tensor of objects. Each object in the tensor contains six values. i.e., x, y, w, h, confidence, label. Here, (x, y) is the center of the detected box, and w, h are the width and height of the box, whereas confidence reflects how likely the box contains an object and how accurate is the bounding box, and finally label is the object that is detected.

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

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.

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. gTTS sends text to Google's servers to generate a speech file, which is then returned. The speech generated by pyttsx3 is spoken comparatively faster than gTTS. The speech generated by both the libraries are 100% accurate.

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 frequently visiting homes, will give better safety and recognition for visually impaired people.

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.

[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] Annapoorani, A., Kumar, N.S., Vidhya, V., 2021. Blind-sight: Object detection with voice feedback .

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] Chen, X., Xu, J., Yu, Z., 2018. A 68-mw 2.2 tops/w low bit width and multiplierless dcnn object detection processor for visually impaired people. IEEE Transactions on Circuits and Systems for Video Technology 29, 3444–3453.

[10] 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.

[11] Fang, W., Wang, L., Ren, P., 2019. Tinier-yolo: A real-time object detection method for constrained environments. IEEE Access 8, 1935–1944.

[12] 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.

[13] Jhinkwan, P., Ingale, V., Chaturvedi, S., 2019. Object detection using convolution neural networks, in: Proceedings of International Conference on Communication and Information Processing (ICCIP).

[14] 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.

[15] Li, Y., Li, S., Du, H., Chen, L., Zhang, D., Li, Y., 2020. Yolo-acn: Focusing on small target and occluded object detection. IEEE Access 8, 227288–227303.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] Rahman, M.A., Sadi, M.S., 2021. Iot enabled automated object recognition for the visually impaired. Computer Methods and Programs in Biomedicine Update , 100015.

[27] 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.

[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] Skalski, P., 2019. Make Sense available at, https://www.makesense.ai/. URL: https://www.makesense.ai/. [Online; accessed 6-November-2021].

[30] Sun, M., Ding, P., Song, J., Song, M., Wang, L., 2019. "watch your step": Precise obstacle detection and navigation for mobile users through their mobile service. IEEE Access 7, 66731–66738.

[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] Ye, C., Qian, X., 2017. 3-d object recognition of a robotic navigation aid for the visually impaired. IEEE Transactions on Neural Systems and Rehabilitation Engineering 26, 441–450.

[36] 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.