THIRD EYE: OBJECT RECOGNITION AND SPEECH GENERATION FOR VISUALLY IMPAIRED

A Project Report Submitted

in Partial Fulfilment of the Requirements

for the Degree of

Bachelor of Technology

in

Computer Science and Engineering

by

Yarlagadda Sai Bhavadeesh (Roll No. 2018BCS0082)

Peddi Shwejan (Roll No. 2018BCS0047)

Allu Harsha Vardhan (Roll No. 2017BCS0005)

Lavanya S (Roll No. 2017BCS0034)

to

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY

KOTTAYAM-686635, INDIA

November 2021

DECLARATION

We, **Yarlagadda Sai Bhavadeesh** (**Roll No: 2018BCS0082**), **Peddi Shwejan** (**Roll No: 2018BCS0047**), **Allu Harsha Vardhan** (**Roll No: 2017BCS0005**), **Lavanya S** (**Roll No: 2017BCS0034**), hereby declare that, the report entitled **"Third Eye: Object Recognition And Speech Generation For Visually Impaired"** submitted to Indian Institute of Information Technology Kottayam towards partial requirement of **Bachelor of Technology** in **Computer Science and Engineering** is an original work carried out by us under the supervision of **Dr. Koppala Guravaiah** and has not formed the basis for the award of any degree or diploma, in this or any other institution or university. We have sincerely tried to uphold the academic ethics and honesty. Whenever an external information or statement or result is used then, that have been duly acknowledged and cited.

Kottayam-686635 **Yarlagadda Sai Bhavadeesh - 2018BCS0082**

November 2021 **Peddi Shwejan - 2018BCS0047**

Allu Harsha Vardhan - 2017BCS0005

Lavanya S - 2017BCS0034

CERTIFICATE

This is to certify that the work contained in this project report entitled **"Third Eye: Object Recognition And Speech Generation For Visually Impaired"** submitted by **Yarlagadda Sai Bhavadeesh** (**Roll No: 2018BCS0082**), **Peddi Shwejan** (**Roll No: 2018BCS0047**), **Allu Harsha Vardhan** (**Roll No: 2017BCS0005**), **Lavanya S** (**Roll No: 2017BCS0034**) to Indian Institute of Information Technology Kottayam towards partial requirement of **Bachelor of Technology** in **Computer Science and Engineering** has been carried out by them under my supervision and that it has not been submitted elsewhere for the award of any degree.

Kottayam-686635 (Dr. Koppala Guravaiah)

November 2021 Project Supervisor

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. Your eyesight is one of your most important senses: 80% of what we perceive comes through our sense of sight [10]. 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.

We aim 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. The thesis employs training on YOLO (You Only Look Once) v5, Convolutional Neural Network (CNN) model for object detection. YOLO v5 is trained on custom dataset of 15 objects, along with MS COCO 2017 Dataset of 80 objects (95 objects overall). In future, the output of model is converted to audio format and is presented to visually impaired.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Visually Impaired face a lot of difficulties in their daily lives. According to World Health Organization (WHO), at least 2.2 billion people have a near or distant vision impairment. Out of them, 49.1 million people are completely visually impaired. Yet the growth of population is making a substantial improvement in the number of people affected. Notable inter-regional and gender inequalities exist which highlights the need to scale up vision impairment alleviation efforts at all levels.

Visually impaired always need the help of either a stick or a person. Young children with early-onset severe vision impairment can experience limited language, emotional, social, and cognitive development, with lifelong consequences. Vision impairment critically impacts the quality of life among the adult population. In the case of older adults, vision impairment can contribute to social isolation, difficulty walking, a higher risk of falls and fractures, and a greater likelihood of early entry into nursing or care homes.

Hence, we took up this project to help visually impaired people to recognize their surroundings. In recent years, deep learning has become a more popular technique for solving these problems of identifying objects. The deep learning systems achieve high accuracy rates at a lower cost. Many Convolutional Neural Network (CNN) methods like Single Shot Detector (SSD) and You Only Look Once (YOLO) are used to solve detection and recognition issues. There are other architectures such as Faster R-CNN and Mask R-CNN [37]. In this project, we used the YOLO v5 algorithm. Implemented a custom-made dataset including MS COCO 2017 dataset to attain good accuracy, and performance. After detecting and recognizing the objects, we are planning to generate speech, for the recognized objects. This can be achieved by using Recurrent Neural Network Techniques (RNN).

YOLO is an abbreviation for the term "You Only Look Once". YOLO algorithm detects and recognizes various objects in a picture (in real-time). YOLO uses convolution neural networks to provide real-time object detection. YOLO v5 is faster, more accurate, and light-weight compared to other versions of YOLO. It has been used in various applications such as autonomous car driving, etc. YOLO v5 is one of the best available models for Object Detection at the moment.

Chapter 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.[28] 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 and it sends the image label. Elmannai, Wafa M and Khaled M. Elleithy. [12] proposed a system where two camera sensors are used for object detection, which is processed using computer vision methods. The remote server handles the image processing. Based on the depth of the image, we can approximately measure the distance between the obstacle and the Visually Impaired person. The Oriented FAST and Rotated BRIEF (ORB) and KNN are used for object detection. Ye, Cang and Xiangfei Qian. [36] 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 the captured image is preprocessed using Convolutional Neural Network, then classification and identification are done using Support Vector Machine Algorithm. Here they used Marshfield Clinic Dataset. Gianani, Sejal et al. [14] came up with a system where the image is captured through a camera device for the input and preprocessed using OpenCV. They have combined both the Single Shot Detector (SSD) framework and the MobileNet architecture to arrive at a fast and efficient deep learning-based method for object detection. Nishajith, A et al. [21] 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. [24] 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. Further, input from the ultrasonic sensor will be utilized to confirm object detection output. Additionally, an IR sensor will detect small objects near feet. Tosun, Selman and Enis Karaarslan. [33] proposed a system where the image is captured using the android platform and preprocessing is done using OpenCV and Tiny YOLO which is implemented using Tensorflow is used for object detection which gives the audio output.

Wong, Yan Chiew et al. [34] In 2019, an object detection system for visually disabled people based on CNN in real-time has been proposed. To reduce the complex load, regional suggestions from the edges of each image map were generated using the control box algorithm. Then, the suggestions passed through a well-configured CaffeNet model. The object group was filmed by a webcam in real-time and the image feature was removed. Next, a sound-based detector was developed to detect the sight of visually impaired people. Nasreen, Jawaid et al. [20]Proposed a system that can be used to guide visually impaired people for object detection. The developed system takes an image from the back camera and loads it into a website and it passes the image to the server, on the server-side YOLO model is used to detect the objects. Pardasani, Arjun et al. [22]They 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. citerahman2019assistive This paper presents an object detection model using the YOLO algorithm for visually impaired people. MTCNN is utilized for the building model. For Object identification and Facial Recognition, YOLO Algorithm and MTCNN Networking are used, respectively. Shah, Samkit et al. [29]In this paper they 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. [15] They proposed a system that uses a convolutional network combined with fully connected layers. They have used the CIFAR-100 dataset, The model is trained with a back-propagation algorithm to detect the objects in the image. Chen, Xiaobai et al. [11] designed an automatic DCNN quantization algorithm to significantly reduce data range up to 4 or 5 bits, reducing hardware costs by more than 68% compared to the 16-bit fixpoint model with irreversible accuracy loss. Sun, Minghui et al. [31] proposed a system using Google Tango, a built-in infrared (IR) sensor to collect data.

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. [13] 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. [17] In the same year, 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. [37] 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. [16] 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. [27] In 2021, proposed an IoT-enabled

Automated Object Recognition where they used SSD Model, SIFT, and MS COCO dataset. Balachandar, Santhosh et al. [5] They proposed a scheme where a multi-view object tracking (MVOT) system is used in this proposed system to address multiple cameras monitoring recording videos. And by combining the knowledge contained in the videos, a powerful and accurate framework is developed. Each segmented group of objects in one view is mapped to the corresponding group in another view using the Yolo V3 algorithm These agreeing sets corresponded to blob gatherings, Which allow data to be exchanged between cameras. These images are transformed into voice output after they are captured by the camera. Mansi Mabendru and Sanjay Kumar Dubey. [19] In this paper a system is developed using two different algorithms i.e. Yolo and Yolo_v3 and tested under the same criteria to measure the accuracy and performance. In the YOLO Tensor flow, the SSD Mobile Net model and in Yolo_v3 Darknet model are used. To get the audio Feedback gTTS (Google Text to Speech), the python library is used to convert statements into audio speech. To play the audio pygame python module is used. Kanchan Patil et al. [25] They proposed a wearable device with a Virtual assistant system for the visually impaired person, Total of five components they merged into one system in this project. The navigation through these components is possible through hardware buttons and voice-over commands given by the user. There are many deep learning methodologies and core libraries of python language used for programming. Mohana Priya et al. [4] In this paper a voice-based image caption generation is a task that involves the NLP (natural language processing). The combination of CNN and LSTM is considered the best solution in this project; the main target of this proposed research work is to obtain the perfect caption for an image. After obtaining the description, it will be converted into text and the text into a voice. Annapoorani et al. [3] You Only Look Once (YOLO) a Real-Time Object Detection is deployed in this paper, Image classification techniques are used to identify the features of the image and Indian currency recognition module is developed to identify the denominations. The text description of the recognized object will be sent to the Google Text-to-Speech API using the gTTS package. Sandeep Pandasupuleti et al. [23] In this paper they proposed Voice Translation and Image Recognition using VCC, LSTM, and Flickr_8k dataset.

Table 2.1: Trends & Technologies discussed in literature

Paper Title	**Authors**	**Methods**	**Pros & Cons**
Proposed System on Object Detection for Visually Impaired People. [28]	Rajwani, Roshan, Dinesh Purswani, Paresh Kalinani, Deesha Ramchandani, and Indu Dokare	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.

Continued on next page

Table 2.1: *Trends & Technologies discussed in literature ... Contd.*

Paper Title	**Authors**	**Methods**	**Pros & Cons**
A Highly Accurate and Reliable Data Fusion Framework for Guiding the Visually Impaired. [12]	Elmannai, Wafa M., and Khaled M. Elleithy	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.
3-D Object Recognition of a Robotic Navigation Aid for the Visually Impaired [36]	Ye, Cang, and Xiangfei Qian	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]	Bashiri, Fereshteh S, Eric LaRose, Jonathan C. Badger, Roshan M. D'Souza, Zeyun Yu, and Peggy Peissig.	Marshfield Clinic Dataset, Google Glass Device, CNN Model, Support Vector Machine Algorithm	Limited number of objects (ex: doors, stairs, signs etc.,) Accuracy over 98%

Continued on next page

Table 2.1: *Trends & Technologies discussed in literature ... Contd.*

Paper Title	**Authors**	**Methods**	**Pros & Cons**
JUVO - An Aid for the Visually Impaired [14]	Gianani, Sejal, Abhishek Mehta, Twinkle Motwani, and Rohan Shende	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. [21]	Nishajith, A., J. Nivedha, Shilpa S. Nair, and J. Mohammed Shaffi.	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 [24]	Patel, Charmi T.,Vaidehi J. Mistry, Laxmi S. Desai, and Yogesh K. Meghrajani.	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

Table 2.1: *Trends & Technologies discussed in literature ... Contd.*

Paper Title	**Authors**	**Methods**	**Pros & Cons**
Real-Time Object Detection Application for Visually Impaired People: Third Eye. [33]	Tosun, Selman, and Enis Karaarslan	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. [34]	Y.C. Wong, J.A. Lai, S.S.S. Ranjit, A.R. Syafeeza, N. A. Hamid	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. [20]	Jawaid nasrren, warsi, Arif, Asad ali shaikh, Yahya Muhammad, Monaisha abdullah.	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.

Continued on next page

Table 2.1: *Trends & Technologies discussed in literature ... Contd.*

Paper Title	**Authors**	**Methods**	**Pros & Cons**
Smart Assistive Navigation Devices for Visually Impaired People. [22]	Arjun Pardasani, Prithviraj N Indi, Sashwata Banerjee, Aditya Kamal, Vaibhav Garg	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 [26]	FerdousiRahman, IsratJahanRitun, NafisaFarhin, JiaUddin	Open CV, YOLO algorithm, Deep learning	The object detection process achieved 6-7 FPS processing with an accuracy rate of 63-80%

Continued on next page

Table 2.1: *Trends & Technologies discussed in literature … Contd.*

Paper Title	**Authors**	**Methods**	**Pros & Cons**
CNN based Auto-Assistance System as a Boon for Directing Visually Impaired Person. [29]	Samkit Shah , Jayraj Bandariya , Garima Jain , Mayur Ghevariya , Sarosh Dastoor	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 [15]	Piyush Jhinkwan , Vaishali Ingale , Shubham Chaturvedi	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

Table 2.1: *Trends & Technologies discussed in literature ... Contd.*

Paper Title	**Authors**	**Methods**	**Pros & Cons**
A 68 mw 2.2 Tops/w low bit-width and multiplierless DCNN object detection processor for visually impaired people. [11]	Xiaobai Chen, Jinglong Xu	Deep convolutional network, low-bit, multiplierless	reducing hardware cost by over 68% compared to the 16 bit fixpoint model with negligible accuracy loss
"Watch Your Step": Precise Obstacle Detection and Navigation for Mobile Users Through Their Mobile Service [31]	MINGHUI SUN PENGCHENG DING , JIAGENG SON , MIAO SONG5 , AND LIMIN WANG	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. [35]	Qiwei Xu, Runzi Lin, Han Yue, Hong Huang, Yun Yang, Zhigang Yao	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%.

Continued on next page

Table 2.1: *Trends & Technologies discussed in literature ... Contd.*

Paper Title	**Authors**	**Methods**	**Pros & Cons**
Tinier-YOLO: A Real-Time Object Detection Method for Constrained Environments. [13]	Wei Fang, Lin Wang, Peiming Ren	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. [17]	Yongjun Li, Shasha Li, Haohao Du, Lijia Chen, Dongming Zhang, Yao Li	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. [16]	Rashika Joshi, Meenakshi Tripati, Amit Kumar, Manoj Singh Gaur.	MobileNetSSD (SSD - Single Shot-Detector), PASCAL VOC 2007.	Got pretty good accuracy, but the dataset is small, not sufficient. Only for embedded systems.

Continued on next page

Table 2.1: *Trends & Technologies discussed in literature ... Contd.*

Paper Title	**Authors**	**Methods**	**Pros & Cons**
An Evaluation of RetinaNet on Indoor Object Detection for Blind and Visually Impaired Persons Assistance Navigation. [2]	Mouna Afif, Riadh Ayachi, Yahia Said, Edwige Pissaloux, Mohamed Atri	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]	Mouna Afif, Riadh Ayachi, Edwige Pissaloux, Yahia Said, Mohamed Atri	YOLOv3, 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.

Continued on next page

Table 2.1: *Trends & Technologies discussed in literature ... Contd.*

Paper Title	**Authors**	**Methods**	**Pros & Cons**
Robot Eye: Automatic Object Detection and Recognition Using Deep Attention Network to Assist Blind People. [37]	Ervin Yohannes, Paul Lin, Chih-Yang Lin, Timothy K. Shih	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]	Swapnil Bhole, Aniket Dhok	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. [27]	Md. Atikur Rahman , Muhammad Sheikh Sadi	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

Table 2.1: *Trends & Technologies discussed in literature ... Contd.*

Paper Title	**Authors**	**Methods**	**Pros & Cons**
Deep Learning Technique Based Visually Impaired People Using YOLO V3 Framework Mechanism. [5]	Balachandar, Santhosh, Suriyakrishna, Vigensh, Usharani, Manju Bala	Yolov3,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 [19]	Mansi Mahendru, Sanjay Kumar Dubey	Tensor flow, SSD, Yolo, Yolo_v3, gTTS, Deep Learning	Yolo accuracy is 78.99 and yolov3 92.89% (seems to be better compare to (yolo)
Guidance System for Visually Impaired People. [25]	Kanchan Patil , Avinash Kharat, Pratik Chaudhary , Shrikant Bidgar , Rushikesh Gavhane	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.

Continued on next page

Table 2.1: *Trends & Technologies discussed in literature ... Contd.*

Paper Title	**Authors**	**Methods**	**Pros & Cons**
Building A Voice Based Image Caption Generator with Deep Learning. [4]	Mohana priya R, Dr.Maria Anu, Divya	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]	A. Annapoorani, Nerosha Senthil Kumar, Dr. V. Vidhya	YOLO, COCO Dataset, gTTS	Live object recognition system cannot perform future learning which is a demerit.
Image Recognition and Voice Translation for Visually Impaired. [23]	Sandeep Pasupuleti, Lahari Dadi, Manikumar Gadi, R. Krishnaveni	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

Chapter 3

Proposed Work

YOLO is an algorithm that uses neural networks to provide real-time object detection. This algorithm is popular because of its speed and accuracy. It has been used in various applications to detect traffic signals, people, parking meters, and animals.

YOLO is an algorithm based on regression, instead of selecting the interesting part of an image, it predicts classes and bounding boxes for the whole image in one run of the algorithm. Ultimately, we aim to predict a class of an object and the bounding box specifying object location.

3.1 Methods

The project uses YOLO algorithm that provides real-time object detection using neural networks and YOLO has different versions.

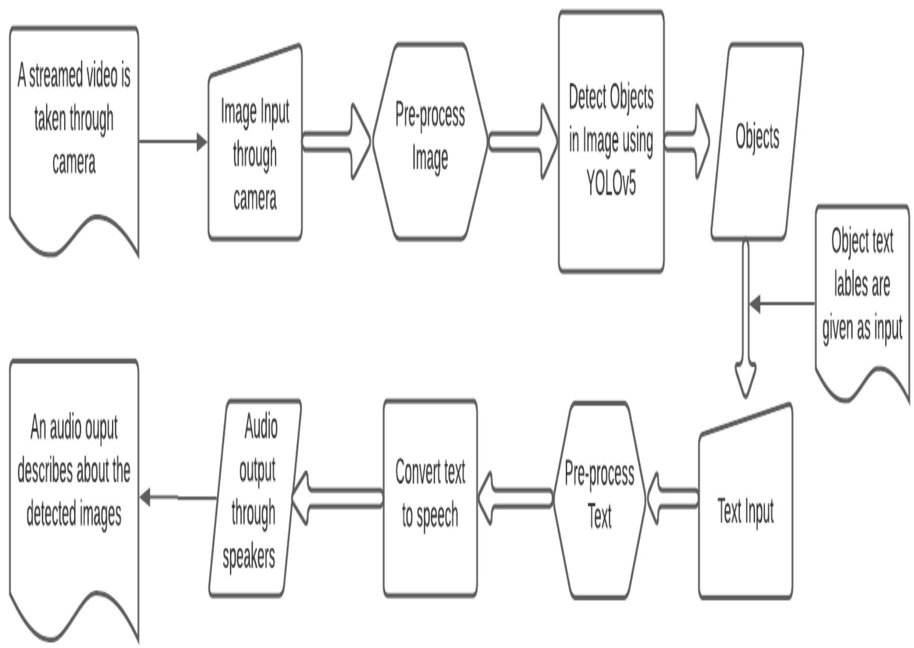

Figure 3.1: Schematic diagram of proposed system

3.1.1 YOLO

YOLO first takes an input image, the framework then divides the input image into grids (say a 3 X 3 grid). Image classification and localization are applied on each grid. YOLO then predicts the bounding boxes and their corresponding class probabilities for objects.

We will divide each image into different grids. For example we divide an image into 3 x 3 grids and there are a total of 3 classes which we want the objects to be classified into. Let's say the classes are Pedestrian, Car, and Motorcycle respectively. So, for each grid cell, the label y will be an eight-dimensional vector: pc defines whether an object is present in the grid or not (it is the probability) bx, by, bh, bw specify the bounding box if there is an object c1, c2, c3 represent the classes. So, if the object is a car, c2 will be 1 and c1 & c3 will be 0, and so on. We will run both forward and backward propagation to train our model.

$$y = \begin{bmatrix} pc \\ bx \\ by \\ bh \\ bw \\ c1 \\ c2 \\ c3 \end{bmatrix}$$

Figure 3.2: Eight Dimensional Vector

3.1.2 YOLO v1

YOLOv1 is a single-stage object detection model. Object detection is framed as a regression problem to spatially separated bounding boxes and associated class probabilities. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. Since the whole detection pipeline is a single network, it can be optimized end-to-end directly on detection performance.

Limitations

- YOLO v1 has difficulties in detecting small objects that appear in groups.

- YOLO v1 has difficulties in detecting objects having unusual aspect ratios.
- YOLO v1 makes more localization errors compared to Fast R-CNN.

3.1.3 YOLO v2

The major improvements of this version are better , faster and more advanced to meet the faster R-CNN which is also an object detection algorithm which uses a Region Proposal Network to identify the objects from the image input and SSD(Single Shot Multibox Detector).

Improvements

- Batch Normalization: it normalizes the input layer by altering slightly and scaling the activations. mAP increased by 2%.
- Higher Resolution Classifier: Input changed from 224*224 to 448*448 mAP increased by 4%.
- Anchor Boxes: are designed to detect objects in the same grid.
- Fine Grained Features: Divides the image into 13*13 grid cells which helps identifying small objects, unlike V1.
- Multi Scale Training: Model is trained on different sizes of objects for the same images.
- Darknet - 19: YOLO v2 uses Darknet 19 architecture with 19 convolutional layers and 5 max-pooling layers and a softmax layer for classification objects. Darknet is a neural network framework written in C

language and CUDA. It's really fast in object detection which is very important for predicting in real-time.

- Results: At 67 FPS, YOLOv2 can give an mAP of 76.8 while at 40 FPS the detector gives an accuracy of 78.6 mAP, better than the state-of-the-model such as Faster R-CNN and SSD while running significantly faster than those models.

3.1.4 YOLO v3

The previous version has been improved for an incremental improvement which is now called YOLO v3. As many object detection algorithms are been there for a while now the competition is all about how accurate and quickly objects are detected. YOLO v3 has all we need for object detection in real-time with accurately and classifying the objects.

Improvements

- Bounding Box Predictions: Uses Logistic Regression to predict the objectiveness score.
- Class Predictions: Uses Logistic classifiers instead of softmax, by doing in so we can have multi- label classification.
- Feature Pyramid Networks.
- Darknet-53 Architecture: has 53 convolutional layers.

3.1.5 YOLO v4

YOLOv4's architecture is composed of CSPDarknet53 as a backbone, spatial pyramid pooling additional module, PANet path-aggregation neck and YOLOv3 head. CSPDarknet53 is a novel backbone that can enhance the learning capability of CNN. The spatial pyramid pooling block is added over CSPDarknet53 to increase the receptive field and separate out the most significant context features. The PANet is used as the method for parameter aggregation for different detector levels instead of FPN used in YOLO v3.

Improvements

- YOLOv4 is twice as fast as EfficientDet (competitive recognition model) with comparable performance.
- YOLO v4 is also based on the Darknet and has obtained an AP value of 43.5 percent on the COCO dataset along with a real-time speed of 65 FPS on the Tesla V100, beating the fastest and most accurate detectors in terms of both speed and accuracy.
- In addition, AP (Average Precision) and FPS (Frames Per Second) increased by 10% and 12% compared to YOLOv3

3.1.6 YOLO v5

So, it said to be that YOLO v5 is extremely fast and lightweight than YOLO v4, while the accuracy is on par with the YOLO v4 benchmark.

YOLO V5 is written in Pytorch framework.

Pytotch inferences are very fast that before releasing YOLOv5, many other AI practitioners often translate the YOLOv3 and YOLOv4 weights into ultralytics Pytorch weight.

Improvements

- YOLO v5 is different from all other prior releases, as this is a PyTorch implementation rather than a fork from the original Darknet.
- Same as YOLO v4, the YOLO v5 has a CSP backbone and PANet neck.
- The major improvements include mosaic data augmentation and auto-learning bounding box anchors.

3.2 Why YOLO v5

YOLO v5 is nearly 90 percent smaller than YOLO v4. So, it said to be that YOLO v5 is extremely fast and lightweight than YOLO v4, while the accuracy is on par with the YOLO v4 benchmark.So we decided to use YOLO V5.

3.3 Dataset

Training, Validation and Testing of proposed model YOLO v5 are done on a custom prepared dataset combined with MS COCO 2017 Dataset [18]. 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 images we considered for doing image detection is 5000.

3.3.1 Annotation tool

We used makesense.ai [30] a data annotation tool to annotate our new dataset. 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.

3.3.2 YOLO format

To train & validate on YOLO algorithm, we need a specific format of dataset, As shown in figure 3.1. In the images folder, we further need to divide it into 3 different folders namely, train, val, test, and save respective images in those folders. similarly for labels folder, here all the labels will be text files. And finally, we need to specify the paths of all the images in respective folders, in their respective text files (train.txt, val.txt, test.txt).

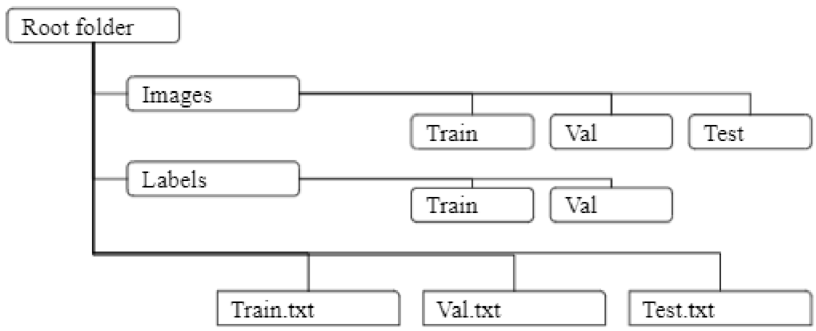

Figure 3.3: Dataset folder structure

3.4 Algorithm

3.4.1 YOLO v4

YOLO stands for You Only Look Once. It's an object detection model used in deep learning use cases. YOLO belongs to the family of One-Stage Detectors (You only look once - one-stage detection). One-stage detection (also referred to as one-shot detection) is that you only look at the image once. YOLO v4 [9] claims to have state-of-the-art accuracy while maintaining a high processing frame rate. It achieves an accuracy of 43.5% AP for the MS COCO with an approximately 65 FPS inference speed on Tesla V100.

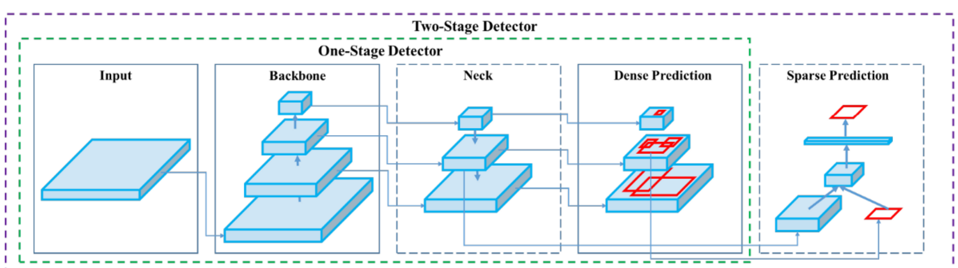

Figure 3.4: YOLO v4 Architecture

In object detection, high accuracy is not the only holy grail anymore. The four parent blocks, after the input image: Backbone (Dense Block & DenseNet, CSP, (CSPDarknet53); Neck (FPN, SPP); Head (Dense Prediction)-used in one-stage-detection algorithms such as YOLO, SSD, etc; Sparse Prediction-used in two-stage-detection algorithms such as Faster-R-CNN, etc (not in YOLOv4).

3.4.2 YOLO v5

YOLO an acronym for 'You only look once', is an object detection algorithm that focuses on detecting objects in images which divides images into a grid system. Each cell in the grid is responsible for detecting objects within itself. YOLO v5 is one of the best available models for object detection at the moment. The great thing about this Deep Neural Network is that it is very easy to retrain the network on our own custom dataset.

Architecture

The network architecture of YOLO v5 [32]. It consists of three parts: Backbone: CSPDarknet, Neck: PANet, Head: YOLO Layer. The data are first input to CSPDarknet for feature extraction and then fed to PANet for feature fusion. Finally, YOLO Layer outputs detection results (class, score, location, size).

Object Detector will have a backbone for pre-training it and a head to predict classes and bounding boxes. The Backbones can be running on GPU or CPU platforms.The Head can be either one-stage (e.g., YOLO, SSD, RetinaNet) for Dense prediction or two-stage (e.g., Faster R-CNN) for the Sparse prediction object detector. Object detectors have some layers (Neck) to collect feature maps, and it is between the backbone and the Head.

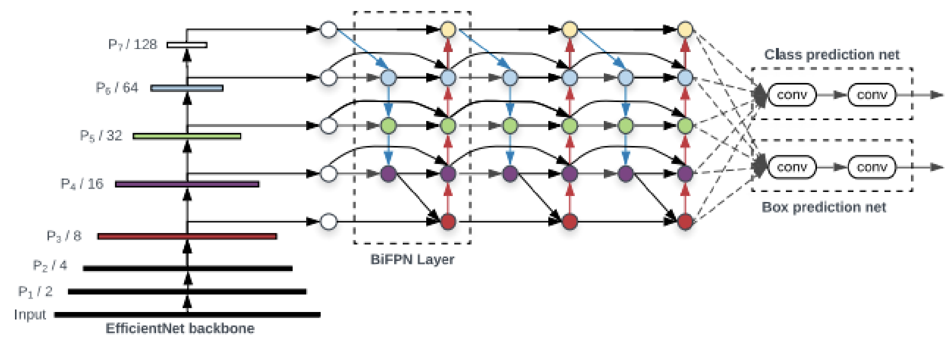

Figure 3.5: EfficientDet architecture

Chapter 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.

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, 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, Box loss, Object loss results for the training set.

Figure 4.1: Training: Class loss vs number of epochs

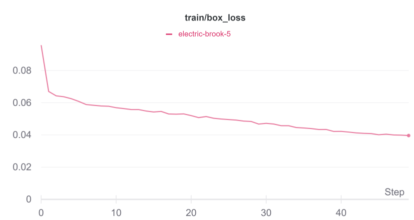

Figure 4.2: Training: Box loss vs number of epochs

Figure 4.3: Training: Object loss vs number of epochs

4.2 Validaton

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, Box loss, Object loss results for the validation set.

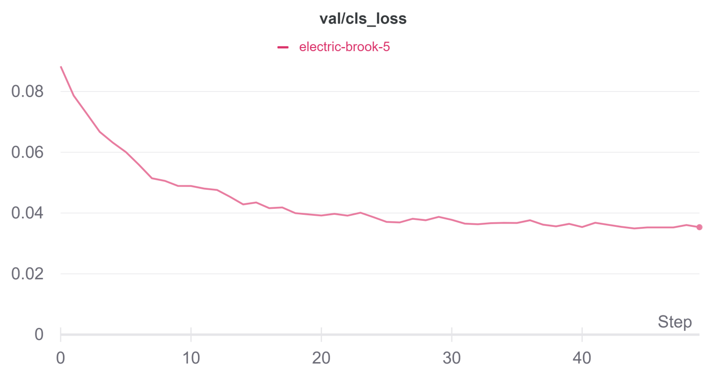

Figure 4.4: Validation: Class loss vs number of epochs

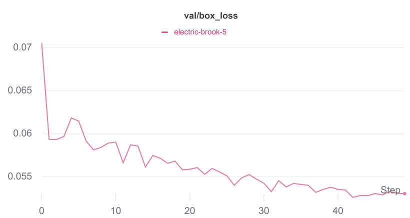

Figure 4.5: Validation: Box loss vs number of epochs

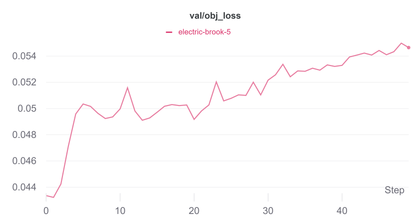

Figure 4.6: Validation: Object loss vs number of epochs

4.3 Evaluation metrics

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

4.3.1 Precision

Precision is one indicator of a machine learning model's performance – the quality of a positive prediction made by the model. Precision refers to the number of true positives divided by the total number of positive predictions (i.e., the number of true positives plus the number of false positives).

4.3.2 Recall

A recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made.

Unlike precision that only comments on the correct positive predictions out of all positive predictions, Recall indicates missed positive predictions.

4.3.3 mean Average Precision (mAP)

The mean Average Precision or mAP score is calculated by taking the mean AP over all classes and/or overall IoU thresholds, Depending on different detection challenges that exist.

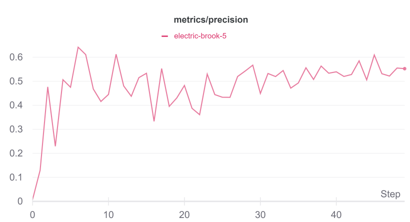

Figure 4.7: Evaluation metric: Precision vs number of epochs

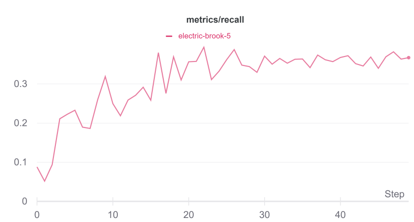

Figure 4.8: Evaluation metric: Recall vs number of epochs

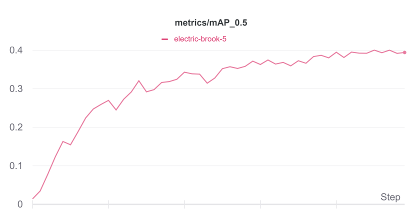

Figure 4.9: Evaluation metric: mAP_0.5 vs number of epochs

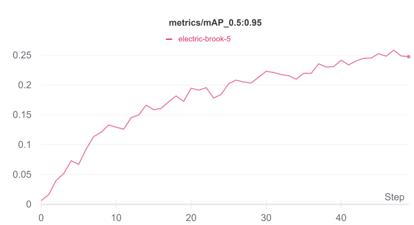

Figure 4.10: Evaluation metric: mAP_0.5:0.95 vs number of epochs

Chapter 5

Conclusion & Future Work

We are able to achieve precision as 0.55, recall as 0.37, and mAP as 0.4, with the proposed model. our model YOLO v5 is able to detect 95 different objects, with high confidence. with this model now we are able to detect objects those are most required for the visually impaired in their daily life.

5.1 Improvising model accuracy

As part of our further work, we try to improve our model accuracy, precision, and recall. And, we also try to make our model detect more objects those are helpful for the visually impaired in their daily life.

5.2 Speech generation

After achieving the convincing metrics, we further move on to convert the detected objects into voice messages, using Recurrent Neural Network (RNN) Architectures. Speech generation gives a better experience to the visually impaired, by letting them know about their surroundings.

Bibliography

[1] Mouna Afif, Riadh Ayachi, Edwige Pissaloux, Yahia Said, and Mohamed Atri. Indoor objects detection and recognition for an ict mobility assistance of visually impaired people. *Multimedia Tools and Applications*, 79(41):31645–31662, 2020.

[2] Mouna Afif, Riadh Ayachi, Yahia Said, Edwige Pissaloux, and Mohamed Atri. An evaluation of retinanet on indoor object detection for blind and visually impaired persons assistance navigation. *Neural Processing Letters*, pages 1–15, 2020.

[3] A Annapoorani, Nerosha Senthil Kumar, and V Vidhya. Blind-sight: Object detection with voice feedback. 2021.

[4] Maria Anu, S Divya, et al. Building a voice based image caption generator with deep learning. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 943–948. IEEE, 2021.

[5] A Balachandar, E Santhosh, A Suriyakrishnan, N Vigensh, S Usharani, and P Manju Bala. Deep learning technique based visually impaired

people using yolo v3 framework mechanism. In *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, pages 134–138. IEEE, 2021.

[6] Fereshteh S Bashiri, Eric LaRose, Jonathan C Badger, Roshan M D'Souza, Zeyun Yu, and Peggy Peissig. Object detection to assist visually impaired people: A deep neural network adventure. In *International Symposium on Visual Computing*, pages 500–510. Springer, 2018.

[7] Swapnil Bhole and Aniket Dhok. Deep learning based object detection and recognition framework for the visually-impaired. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pages 725–728. IEEE, 2020.

[8] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com.

[9] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

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

[11] Xiaobai Chen, Jinglong Xu, and Zhiyi Yu. 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(11):3444–3453, 2018.

[12] Wafa M Elmannai and Khaled M Elleithy. A highly accurate and reliable data fusion framework for guiding the visually impaired. *IEEE Access*, 6:33029–33054, 2018.

[13] Wei Fang, Lin Wang, and Peiming Ren. Tinier-yolo: A real-time object detection method for constrained environments. *IEEE Access*, 8:1935–1944, 2019.

[14] Sejal Gianani, Abhishek Mehta, Twinkle Motwani, and Rohan Shende. Juvo-an aid for the visually impaired. In *2018 International Conference on Smart City and Emerging Technology (ICSCET)*, pages 1–4. IEEE, 2018.

[15] Piyush Jhinkwan, Vaishali Ingale, and Shubham Chaturvedi. Object detection using convolution neural networks. In *Proceedings of International Conference on Communication and Information Processing (ICCIP)*, 2019.

[16] Rashika Joshi, Meenakshi Tripathi, Amit Kumar, and Manoj Singh Gaur. Object recognition and classification system for visually impaired. In *2020 International Conference on Communication and Signal Processing (ICCSP)*, pages 1568–1572. IEEE, 2020.

[17] Yongjun Li, Shasha Li, Haohao Du, Lijia Chen, Dongming Zhang, and Yao Li. Yolo-acn: Focusing on small target and occluded object detection. *IEEE Access*, 8:227288–227303, 2020.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft

coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

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

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

[21] A Nishajith, J Nivedha, Shilpa S Nair, and J Mohammed Shaffi. Smart cap-wearable visual guidance system for blind. In *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 275–278. IEEE, 2018.

[22] Arjun Pardasani, Prithviraj N Indi, Sashwata Banerjee, Aditya Kamal, and Vaibhav Garg. Smart assistive navigation devices for visually impaired people. In *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, pages 725–729. IEEE, 2019.

[23] Sandeep Pasupuleti, Lahari Dadi, Manikumar Gadi, and R Krishnaveni. Image recognition and voice translation for visually impaired. *International Journal of Research in Engineering, Science and Management*, 4(5):18–23, 2021.

[24] Charmi T Patel, Vaidehi J Mistry, Laxmi S Desai, and Yogesh K Meghrajani. Multisensor-based object detection in indoor environment for visually impaired people. In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1–4. IEEE, 2018.

[25] Kanchan Patil, Avinash Kharat, Pratik Chaudhary, Shrikant Bidgar, and Rushikesh Gavhane. Guidance system for visually impaired people. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 988–993. IEEE, 2021.

[26] Ferdousi Rahman, Israt Jahan Ritun, Nafisa Farhin, and Jia Uddin. An assistive model for visually impaired people using yolo and mtcnn. In *Proceedings of the 3rd International Conference on Cryptography, Security and Privacy*, pages 225–230, 2019.

[27] Md Atikur Rahman and Muhammad Sheikh Sadi. Iot enabled automated object recognition for the visually impaired. *Computer Methods and Programs in Biomedicine Update*, page 100015, 2021.

[28] Roshan Rajwani, Dinesh Purswani, Paresh Kalinani, Deesha Ramchandani, and Indu Dokare. Proposed system on object detection for visually impaired people. *International Journal of Information Technology (IJIT)*, 4(1):1–6, 2018.

[29] Samkit Shah, Jayraj Bandariya, Garima Jain, Mayur Ghevariya, and Sarosh Dastoor. 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), pages 235–240. IEEE, 2019.

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

[31] Minghui Sun, Pengcheng Ding, Jiageng Song, Miao Song, and Limin Wang. "watch your step": Precise obstacle detection and navigation for mobile users through their mobile service. *IEEE Access*, 7:66731–66738, 2019.

[32] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[33] Selman Tosun and Enis Karaarslan. Real-time object detection application for visually impaired people: Third eye. In *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, pages 1–6. Ieee, 2018.

[34] Yan Chiew Wong, JA Lai, SSS Ranjit, AR Syafeeza, and NA Hamid. Convolutional neural network for object detection system for blind people. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 11(2):1–6, 2019.

[35] Qiwei Xu, Runzi Lin, Han Yue, Hong Huang, Yun Yang, and Zhigang Yao. Research on small target detection in driving scenarios based on improved yolo network. *IEEE Access*, 8:27574–27583, 2020.

[36] Cang Ye and Xiangfei Qian. 3-d object recognition of a robotic navigation aid for the visually impaired. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(2):441–450, 2017.

[37] Ervin Yohannes, Paul Lin, Chih-Yang Lin, and Timothy K Shih. Robot eye: Automatic object detection and recognition using deep attention network to assist blind people. In *2020 International Conference on Pervasive Artificial Intelligence (ICPAI)*, pages 152–157. IEEE, 2020.