International Journal of Engineering Research and Science & Technology

SMART OBJECT DETECTION AND SPEECH GUIDANCE SYSTEM FOR THE VISUALLY IMPAIRED

S. Karuna Sri[1], T. Anusha[2], P. Sowmya Sri[3], T. Harika Sai[4]

[1]Assistant Professor in Dept. of CSE, Vignan's Institute of Management and Technology for Women , Ghatkesar, Telangana, India
[2,3,4]B. Tech 4rd year Students, CSE, Vignan's Institute of Management and Technology for Women, Ghatkesar, Telangana, India

Abstract:
An intelligent assistive system that provides object detection and voice instructions is proposed to aid the visually impaired in performing daily tasks. The system includes a camera for capturing video feed from which objects are detected using a deep learning technique referred to as YOLO. Spatial positions of objects including left, center, or right and distances between the user and objects are calculated. Detected data is translated into voice instructions using the text-to-speech technology. Stereo sound is employed in order to aid with directionality. The model is trained using data from COCO dataset to provide accuracy in object recognition. The system is relatively light-weight and does not involve the use of expensive equipment thus affordable. It functions in real-time without any delay in response. The system is effective both indoors and outdoors.

Keywords: Object Detection, YOLO Algorithm, Computer Vision, Assistive Technology, Visually Impaired, Speech Guidance, Text-to-Speech (TTS), Real-Time Systems, Spatial Awareness, Deep Learning

I. INTRODUCTION:

Visual disability faces a lot of difficulties especially concerning perception and interaction with the environment around. Conventional assistive devices like white cane and guide dog can help detect obstacles, but they don't provide any information on the nature, location and movement of the objects around. Thanks to recent advancements in AI, computer vision and deep learning, it has become possible to develop more sophisticated systems of assistance that could overcome these problems. In this regard, our work aims at developing an assistive system named Smart Object Detection and Speech Guidance for blind people. Our system uses camera to receive live visual data and processes this data with YOLO (You Only Look Once) algorithm in order to detect objects in the frame and their spatial coordinates, which could be left, center, or right. Further, this data is converted into speech by means of text-to-speech technology. In doing so, user can recognize his environment through hearing and thus increase his perception of surroundings. The system we propose is characterized by lightness, simplicity and low cost and requires neither advanced equipment nor complicated algorithms. Combining the functions of computer vision and speech synthesis ensures an easy-to-use product. In addition, the application was created to operate both indoors and outdoors, giving it flexibility. The entire project is focused on

improving the lives of people with visual disabilities by using advanced technologies.

II. RELATED WORK:

New developments in technology that are aimed at assisting visually impaired people have been directed towards the incorporation of artificial intelligence, computer vision, and deep learning, all of which can be used to improve environmental awareness and navigation.

Many studies have examined real-time object detection solutions using deep learning models such as YOLO, SSD, and Faster R-CNN. Such models can detect multiple objects simultaneously with great precision and speed. Nevertheless, most of the systems only focus on the recognition of the object without offering relevant audio instructions and environmental details needed for navigation purposes [1]. It is already possible to develop software that can help the visually impaired recognize the objects or even texts. Seeing AI and some other OCR software are just a few that can be considered the most popular. These make use of the computer vision and natural language processing to render the visual information into audio information. Despite their success, particularly in helping the visually impaired read and recognize the currency, they still fail to provide constant real-time support as well as provide a more comprehensive view of the environment around them [2]. There is research done on sound-based navigation that employs Text To Speech (TTS) for converting visual inputs into vocal instructions. Such navigation provides information about detected objects via voice output, increasing situational awareness [3]. Yet, most of these technologies lack accurate information regarding position and distance of objects. There are systems that try to incorporate stereo or 3D sound, but they are usually faced with the problem of delay and synchronization, as well as increased computational complexity [4]. Moreover, there are studies that have tried incorporating OCR and scene description techniques. The systems described were able to read signboards and documents, providing even more assistance to users with vision impairment [5]. Yet, the inclusion of such functions as OCR and object detection makes systems even more complex.

III.PROPOSED SYSTEM:

A. Overview of Proposed System:

The proposed Smart Object Detection and Speech Guidance System for Visually Impaired People is a novel assistive technology that makes use of artificial intelligence algorithms to enable greater autonomy among people with limited or no vision. The system takes real-time visual inputs through its camera and uses computer vision and deep learning algorithms to detect objects in the environment. To achieve the objective, the system leverages the efficiency of YOLO (You Only Look Once) algorithm which makes it capable of detecting multiple objects simultaneously. After detecting objects, the system analyzes the different aspects of the objects like type, position, and distance from each other. Subsequently, information about detected objects is used to generate audio inputs using the text-to-speech feature. In essence, the proposed solution enables users to “listen” to their environment and decide on their actions. What distinguishes the suggested assistive tool from conventional devices is that it is affordable and easy to implement without requiring

costly hardware resources.

System Architecture:
The system architecture consists of the following modules:

- Input Layer
- Preprocessing Layer
- Object Detection Layer

- Analysis Layer
- Voice Generation Layer
- Voice Output Layer
- Cotinuous Feedback Mechanism

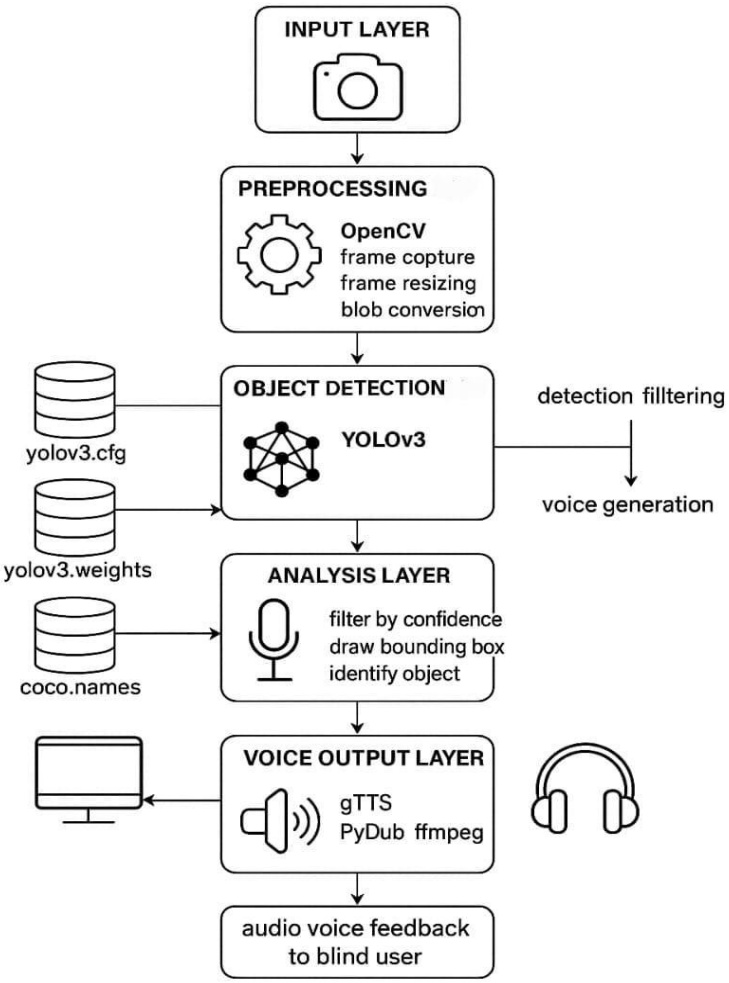


Figure 1: System Architecture

Input Layer: The input layer is meant to collect the live visual data from the environment. In this case, the live frames are collected using a webcam, and these contain different objects that are in proximity to the user.

These live frames are seen as inputs to the system.

Preprocessing Layer: Here, captured frames are preprocessed using the OpenCV library. Preprocessing involves capturing frames, adjusting the size of the frame as per the input required by the model, and converting the frame to the blob form.

Object Detection Layer: This is the key module of our system and it performs the actual object detection process through the YOLOv3 model. The YOLOv3 model uses configuration files (yolov3.cfg), trained weights (yolov3.weights) along with the class labels (coco.names) for object detection and generates bounding boxes, object names along with the confidence levels for the detected obje ct.

Analysis Layer: The analysis layer will analyze the objects detected by vision processing to obtain useful information from the detected objects. The detection will categorize the objects according to the level of confidence, types, and boxes drawn around the objects. Additionally, the analysis layer will determine the spatial position of the objects, whether it is on the left, right, or center.

Voice Generation Layer: In this layer, the object information is converted into text first, which is then converted into speech using various software tools. Software such as gTTS, PyDub, and FFmpeg help in generating the voice messages.

Voice Output Layer: This layer produces audio feedback to the user via the use of speakers or headphones. In other words, this is the output mechanism for the feedback system, providing visual impaired individuals with voice assistance.

Continuous Feedback Mechanism: As mentioned earlier, the whole system works in cycles, in which the latest image frames are captured and analyzed to generate real-time voice feedback.

IV.IMPLEMENTATION:

A. Data Collection:

The model employs open-source data such as COCO (Common Objects in Context) datasets in order to enable it to train the object detection model. Open source datasets have annotated images containing different kinds of objects that range from people, chairs, cars, and various other commonly seen objects. Real-life data collection occurs via the camera when the model is operational to detect objects around the user.

B. Data Preprocessing:

Data preprocessing entails modifying the data so that it can enable the correct prediction of results. The data preprocessing process comprises resizing of the images to specific dimensions, normalization of the pixels, label bounding boxes on objects, and label the objects themselves. There could also be other preprocessing methods such as data augmentation where random flipping, rotations, and scaling of images occur.

C. Model Training:

YOLO (You Only Look Once), a deep learning algorithm used for object detection, is used to detect and classify the objects in the image. In training the model using the collected data, the model will be able to predict the coordinates of the bounding boxes alongside the probability of classes.

D. Prediction:

The YOLO trained algorithm takes up each frame in real time coming from the camera feed and identifies the objects in the frame and gives out the class labels along with bounding boxes. It also computes whether

the detected object is on the left side, right side, or middle position in the frame.

E. Voice Interface:

Information related to the detected object is further conveyed by translating them into voice using Text-To-Speech (TTS) software. Either Pyttsx3 or gTTS library can be used for creating the output audio message. Voice messages such as "Person on the left" and "Table in front" will be created.

F. Result Output:

Output produced by the system will take the form of an audio output. The system will convey live information about the surroundings to the user via audio messages.

V. ALGORITHM:

- **Step 1: Begin**

- **Step 2: Initialization**
  - Load necessary modules (cv2, numpy, yolov8, subprocess, time)
  - Define configuration settings:
  - Confidence threshold

  - Announcement interval
  - Create global variables for announcement tracking

- **Step 3: Load Model and Classes**
  - Load YOLOv8 model
  - Read COCO classes from dataset

- **Step 4: Initialize Video Stream**
  - Capture video stream via OpenCV
  - Verify video stream availability
  - Exit if unavailable

- **Step 5: Capture Frame**
  - Continuously capture frames from webcam stream
  - Skip frame on failure
- **Step 6: Pre-processing**
  - Scale frame to a standard size (e.g., 416 x 416)
  - Prepare frame for efficient object detection

- **Step 7: Object Detection (YOLO)**
  - Pass frame to YOLO model
  - Retrieve bounding box, object label, and confidence score
  - Specify minimum confidence score threshold for object detection results
- **Step 6: Pre-processing**
  - Scale frame to a standard size (e.g., 416 x 416)
  - Prepare frame for efficient object detection

- **Step 7: Object Detection (YOLO)**
  - Pass frame to YOLO model

- Retrieve bounding box, object label, and confidence score
- Specify minimum confidence score threshold for object detection results

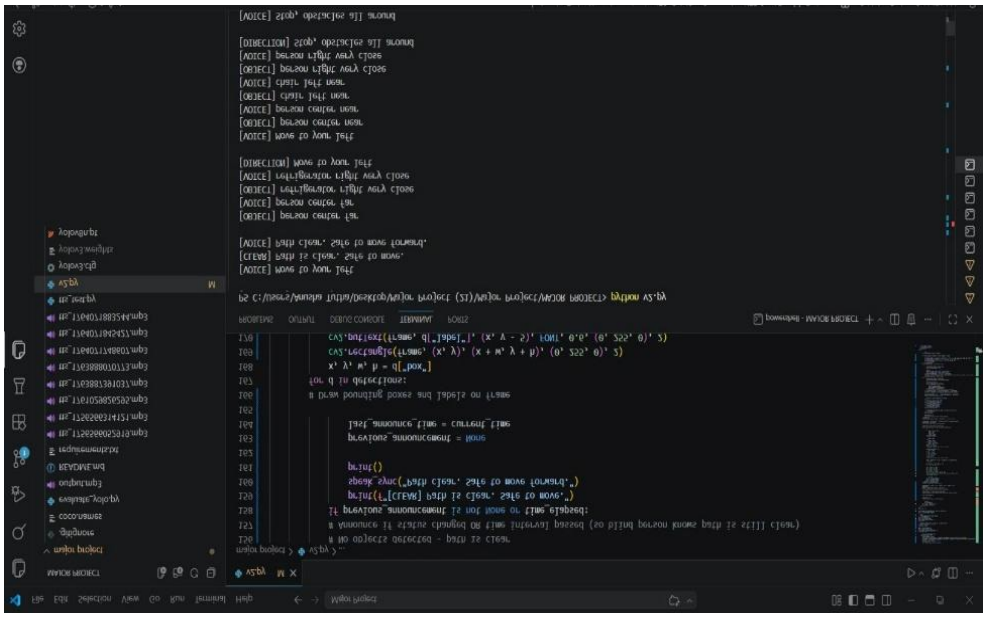
- **Step 8: Coordinate Transformation**
  - Convert bounding box coordinates into frame size
- **Step 9: Feature Extraction**
  - For every object detected:
  - Identify the object
  - Compute the coordinates of its center point (cx, cy)
  - Establish its horizontal alignment (left, middle, or right)
  - Assess its distance (distant, nearby, or very close)

- **Step 10: Environmental Analysis**
  - Monitor directions occupied by objects (left, middle, right)
  - List the detected objects from left to right

- **Step 11: Decision Making (Navigate Support)**
  - If there are no obstacles in the middle direction, “Move forward.”
  - If there are no obstacles in the left direction, “Turn to the left.”
  - If there are no obstacles in the right direction, “Turn to the right.”
  - Otherwise, “Stop, there are obstacles surrounding you.

- **Step 12: Speech Announcement Construction**
  - Create the speech announcement including:
  - Names of the objects, their locations, and distances
  - Directions

VI.RESULTS:



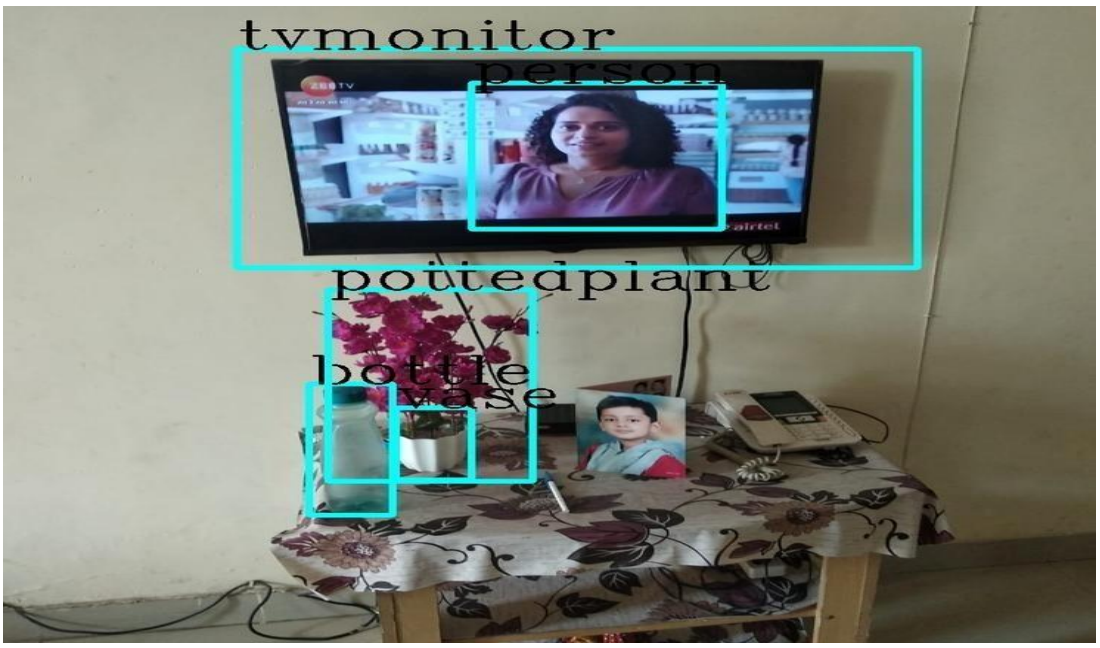


Figure 3: User Input Interface

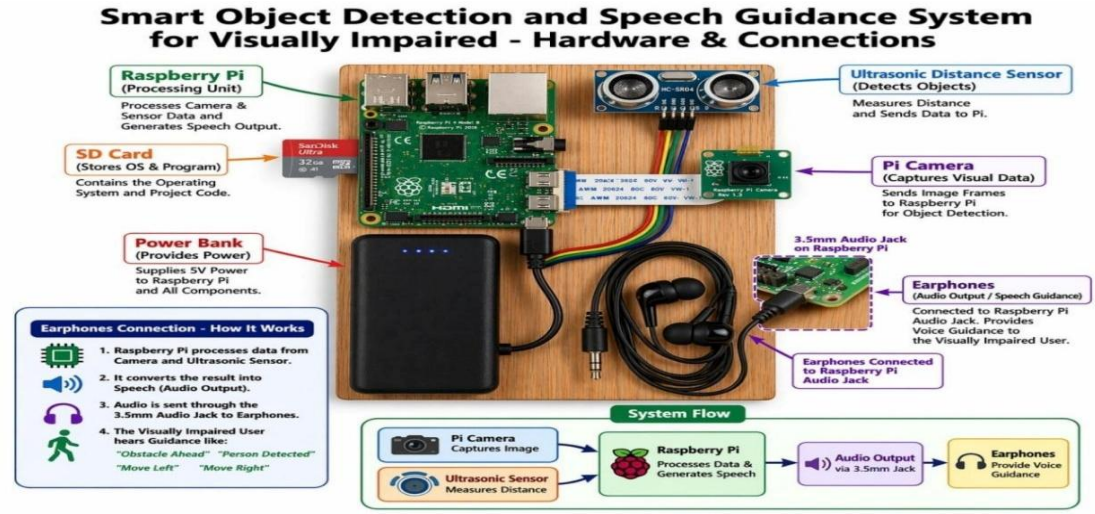
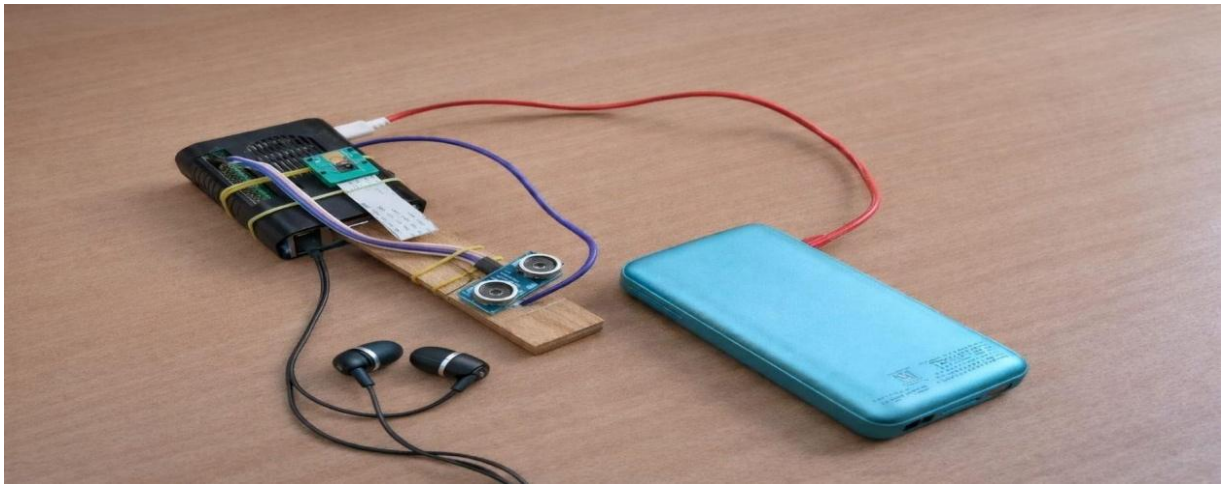


Figure 4: Language Selection Module



VI. CONCLUSION:

The Smart Object Detection and Speech Guidance System enhances autonomy among visually challenged individuals through its integration of object detection technology with speech guidance. Through the utilization of the You Only Look Once (YOLO) algorithm, the system is able to detect various objects and determine their locations within the left, center, and right portions of the visual field. The system offers immediate verbal guidance, allowing individuals to traverse obstacles without difficulties.

VII.REFERENCES:

[1] Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," Computer Vision and Pattern Recognition (CVPR), 2018.

[2] R. Szeliski, Computer Vision: Algorithms and Applications, Springer, 2010.

[3] D. A. Forsyth and J. Ponce, Computer Vision: A Modern Approach, Pearson Education, 2012.

[4] A. Rosebrock, Practical Python and OpenCV, PyImageSearch Publications, 2019.

[5] Goodfellow, I., Bengio, Y., and Courville, A., Deep Learning, MIT Press, 2016.

[6] Gonzalez, R. C., and Woods, R. E., Digital Image Processing, Pearson Education, 2018.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[8] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, 2004.

[9] R. Srikanteswara, M. C. Reddy, M. Himateja and M. K. Mahesh, "Object Detection and Voice Guidance for the Visually Impaired Using a Smart App," in *Recent Advances in Artificial Intelligence and Data Engineering*, Springer, 2022, pp. 133–144.

[10] N. Alzahrani and H. H. Al-Baity, "Object Recognition System for the Visually Impaired: A Deep Learning Approach using Arabic Annotation," *Electronics*, vol. 12, no. 3, p. 541, 2023.

[11] M. Obayya, F. N. Al-Wesabi, W. Bedewi *et al.*, "An intelligent framework for visually impaired people through indoor object detection-based assistive system using YOLO with recurrent neural networks," *Scientific Reports*, vol. 15, 2025.

[12] N. Jawaid, A. Warsi, A. A. Shaikh and M. Yahya, "Object Detection and Narrator for Visually Impaired People," in *Proc. 2019 IEEE 6th Int. Conf. on Engineering Technologies and Applied Sciences (ICETAS)*, 2019.

[13] M. Afif *et al.*, "Third Eye: Object Recognition and Speech Generation for Visually Impaired," *Procedia Computer Science*, vol. 218, pp. 1144–1155, 2023.

[14] D. Ravi Kumar, H. K. Thakkar, S. Merugu, V. K. Gunjan and S. K. Gupta, "Object Detection System for Visually Imppaired Persons Using Smartphone," in *ICDSMLA 2020*, Lecture Notes in Electrical Engineering, Springer, 2022.

[15] A. M. George, A. Ramachandran, M. Ajnas and P. Subeh, "YOLO-Based Object Recognition System for Visually Impaired," *International Journal of Science and Engineering Applications*, vol. 14, no. 1, pp. 34–42, 2025.

[16] P. Boobalan, B. S., M. Sivapriya and R. Sivakumar, "Object Detection with Voice Guidance to Assist Visually Impaired Using YOLOv7," *International Journal for Research in Applied Science & Engineering Technology*, 2023.

[17] M. Noman, V. Stankovic and A. Tawfik, "Portable Offline Indoor Object Recognition System for the Visually Impaired," *Cogent Engineering*, vol. 7, no. 1, 2020.