ISSN: 2093-5374 / E-ISSN: 2093-5382

VocalEyes: A Context-Aware Real-Time Assistive Vision System with Priority-Based Audio Guidance

Dr.P. Renukadevi[1*], Dr.M. Kumaresan[2], Dr.P. Manikandan[3], and N. Shivani[4]

[1*]Assistant Professor, Department of Computer Science and Engineering, Jain (Deemed-To-Be University), Bengaluru, India. pgrenu@gmail.com, https://orcid.org/0009-0001-9533-5860

[2]Associate Professor, Department of Computer Science and Engineering, Jain (Deemed-To-Be University), Bengaluru, India. phdkumaresan@gmail.com, https://orcid.org/0000-0001-6716-7209

[3]Professor, Department of Computer Science and Engineering, Jain (Deemed-To-Be University), Bengaluru, India. mani.p.mk@gmail.com, https://orcid.org/0000-0003-3037-7688

[4]Department of Computer Science and Engineering, Jain Deemed to be University Bengaluru, India. swapna6_pinni@rediffmail.com, https://orcid.org/0009-0002-2096-4053

Received: March 03, 2026; Revised: April 09, 2026; Accepted: May 28, 2026; Published: June 30, 2026

Abstract

Visually impaired people have a significant difficulty in moving, maneuvering, detecting hazards, and taking up visual information on their own. While a large number of assistive technologies have been created, they are mostly devoted to either object identification or text reading, and are not equipped with context-aware decision making. In this paper, VocalEyes, a real-time assistive vision system and a context-aware approach that convert the visual information into intelligent audio information for visually impaired users is introduced. The Plan designed to combine YOLOv8 object detection, distance and direction estimation, priority alert, intelligent silence, optical character recognition (OCR), and offline speech synthesis, all in one architecture. A prioritized decision module sorts the detected objects based on the relevance of the hazard, thus communicating the more important ones in a timely manner, while reducing unnecessary notifications. Furthermore, OCR only operates when there are no high priority hazards and reduces the distraction to the user during navigation. Experimental results show high accuracy (91%), precision (89%), recall (71%), F1-score (79%) and mAP@0.5 (78%) with low response time (less than one second) and a high frame rate (15–25 FPS) in real time. The smart silence feature also optimizes the usability, as it minimizes repetitive announcements and cognitive overloads. The outcomes highlight that VocalEyes is an effective and fully offline assistive solution, which is cost effective, modular and supports the independence of users, the safety of their navigation in the environment and the awareness of their surroundings.

Keywords: Assistive Technology, Visual Impairment, Object Detection, Priority-Based Alerting, OCR, Intelligent Silence, Real- Time Audio Guidance, Context-Aware Systems.

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 2 (June-2026), pp. 663-683. DOI: *10.58346/JOWUA.2026.I2.037*

*Corresponding author: Assistant Professor, Department of Computer Science and Engineering, Jain (Deemed-To-Be University), Bengaluru, India.

1 Introduction

Conventional assistive tools, such as a white cane and guide dog are not sufficient to offer granular contextual awareness, imposing severe constraints on visual impaired people's mobility and interaction with the environment. New techniques such as deep learning, real-time object detection, and optical character recognition (OCR) provide technical solutions to this perceptual gap, but existing paradigms have key usability issues to address (Alahmadi et al., 2023; Ashiq et al., 2022). The majority of the systems use a continuous “detect-and-speak” strategy, which is cognitively demanding since it is expected to talk about all the entities detected (Budrionis et al., 2022). In addition, conventional OCR modules are not discriminative, causing unwanted acoustic noise in safety critical situations (Bai et al., 2018).

To overcome those drawbacks, this paper proposes a novel context-sensitive real-time assistive vision framework, named VocalEyes, which can provide prioritized auditory guidance completely offline (Zhang et al., 2021; Xu et al., 2023). VocalEyes, however, takes a different approach to the black-box or cloud-based solution with a decision-based pipeline that emulates a human guide (Zhang et al., 2021) (Jeong et al., 2025). The system is based on an architecture that combines real-time object detection, BBD distance and direction estimation, a safety-prioritisation engine, an intelligent silence mechanism and safety-gated OCR activation (Katkade et al., 2026).

This work's significant contributions are:

- Integration of custom and COCO based object detection pipeline optimized for Low cost, Offline deployment.
- Real-time decision logic to classify and utter objects according to their significance in the environment. Designing an audio-suppression protocol for context in order to avoid multiple alerts and reduce user's cognitive load. Only using text to speech translation when the immediate environment is deemed safe. Complete operational independence from cloud or internet, and using standard computing hardware.

The rest of this paper will be organized as follows. In Section 2, the literature review of assistive vision system, object detection methods, Optical Character Recognition (OCR) systems and audio navigation systems for visually impaired is presented. A representative case study on the practical implementation of the proposed VocalEyes system for indoor and outdoor environments is discussed in Section 3. The proposed system architecture, object detection model, distance estimation model, direction estimation model, priority-based alert, intelligent silence module and integration of OCR are described in section 4, which includes the system architecture and methodologies and provides the implementation procedures, data preparation, system algorithms, system initialization of parameters and software hardware configuration. The experimental results and discussion, with performance evaluation metrics, confusion matrix analysis, training performance, comparative benchmarking, and ablation studies are provided in Section 5. Finally, in Section 6, the paper ends and future research directions to improve context-aware assistive navigation systems are presented.

2 Literature Review

Traditional assistive methodologies for visually impaired individuals have transitioned from rudimentary proximity alerts using ultrasonic sensors to advanced computer vision frameworks utilizing deep learning models like YOLO and convolutional neural networks (CNNs) for real-time object detection (Alahmadi et al., 2023; Ashiq et al., 2022; Katkade et al., 2026; Mohsenzadegan et al., 2022).

While these contemporary edge-computed and smartphone-based systems successfully identify environmental obstacles and translate them into speech, they generally suffer from an indiscriminate "detect-and-speak" paradigm that yields repetitive, verbose, and unprioritized auditory feedback in dense settings, thereby compounding the user's cognitive load (Alahmadi et al., 2023; Jeong et al., 2025; Kumari & Hammady, 2026). Similarly, optical character recognition (OCR) engines integrated into mobile applications or wearable smart devices effectively transcribe printed materials into audio streams, yet they lack operational integration with environmental awareness, creating severe cognitive distractions by continuously reading text during safety-critical navigational situations (Bai et al., 2018). To mitigate the operational latencies, high hardware expenditures, and data privacy vulnerabilities associated with cloud-dependent or multimodal RGB-D depth architectures, recent paradigms explore compact, quantized models deployed entirely at the edge; however, these frameworks still lack context-driven decision logic or selective audio-suppression mechanisms (Chou et al., 2023; Ashiq et al., 2022; Krishnan et al., 2024).

Analysis of the literature shows that there is a fundamental gap between the capacity of the computer vision system and the human cognitive system, both in terms of the number of objects that can be detected (dozens per second for edge AI, versus continuous language streams for the human auditory system) and in terms of the processing speed of the human brain. Current assistive vision frameworks can be inferred with three primary research gaps based on its operational deficiency. Lack of contextual resource orchestration: current systems and approaches view text reading, object classification, and spatial tracking as separate tasks, with no high-level, supervisory logic to adapt the allocation of secondary resources, such as OCR to the presence of primary resources, such as oncoming vehicles or stairs. Secondly, traditional models are based on distance thresholds or unstructured lists of labels, neglecting to link semantic identity with spatial monitoring to create a united threat map. Third, traditional frameworks fail to record the history of each vocalization, and the lack of contextual memory retention leads to repetitive vocalizations of static environmental features, directly increasing the user mental fatigue. Therefore, an integrated, low-cost and fully offline architecture is required. An efficient assistive system is not supposed to be a continuous data recitation device, but rather a cognitive filter that prioritizes the safety-critical objects, dynamically manages speech suppression, and enforces the safety-gated execution to optimize real-time navigation.

3 Case Studies

This section gives some realistic examples of the successful operation of VocalEyes. The context-aware decision making, priority alerting mechanisms, and intelligent silence mechanisms are shown in each of the case studies.

A. Case Study 1: Indoor Navigation

VocalEyes recognizes people, furniture (like chairs, tables), and doors around it in an indoor setting like an office or classroom by applying the object detection module (Obayya et al., 2025; Wang et al., 2024). The system determines the distance and direction of every object it senses and plays sound cues only for those objects that are of interest (Chou et al., 2023). For example, VocalEyes will alert and tell someone how far and in what direction they are when near. Furniture objects are suppressed until they are on the way to the user. If a screen or a notice board is available, OCR is only possible if there is no obstacle in the way (Obayya et al., 2025). When the spatial configuration is not changed, the system automatically turns off the repeated announcements by intelligent silence (Xu et al., 2023). This case study shows how

the system can be used to provide safe indoor navigation while using minimal unnecessary narration (Chou et al., 2023).

B. Case Study 2: Outdoor Navigation

In outdoor environments, the system focuses on items that are related to safety, such as vehicles and stairs (Katkade et al., 2026; Kumari & Hammady, 2026). VocalEyes activates an alarm and disables OCR when a vehicle drives up on the road. The closer you are to the threat, the more urgent the alert, the closer the better (Chou et al., 2023). Direction awareness gives messages like “vehicle approaching from right” to aid in spatial knowledge (Jeong et al., 2025). This case-study shows the effectiveness of the priority-based alerting in a dynamic outdoor environment.

4 Framework and Methodologies

System Framework

VocalEyes' technical architecture is designed to be modular, localized, and to process real-time visual streams in an offline manner without relying on a cloud service, thus guaranteeing low latency. The execution flow (as shown in figure 1) goes in the sequence of:

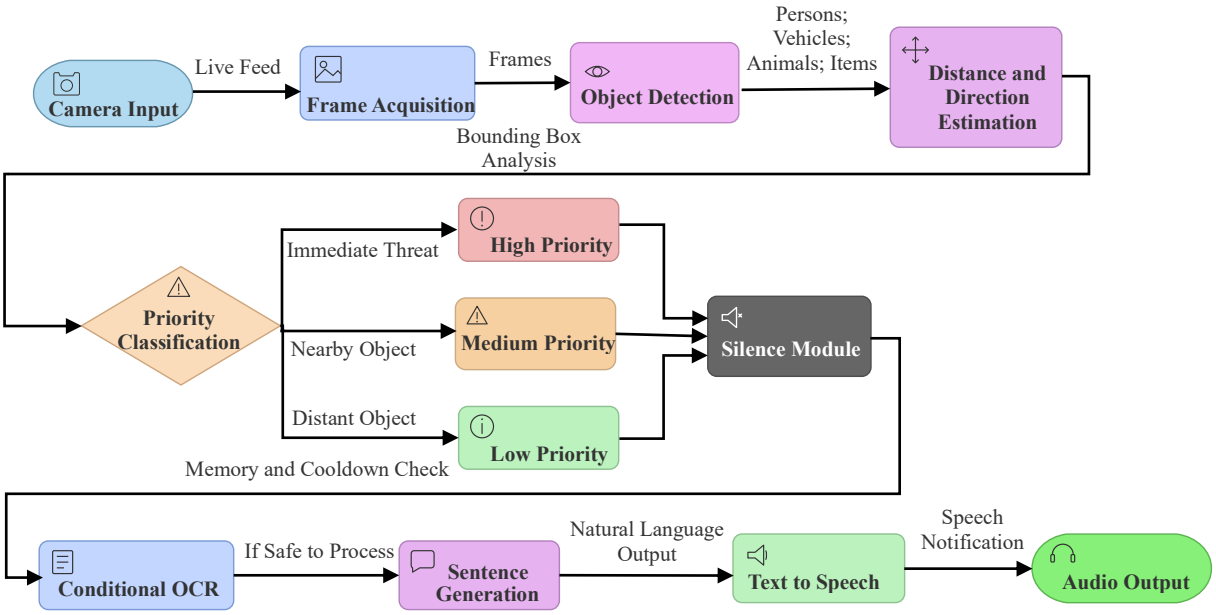


Figure 1: System architecture of the VocalEyes context-aware assistive vision system

The deep learning-based object detection module detects the entities in the environment, including pedestrians, vehicles, and obstacles, from incoming video frames (Ramadhan, 2018); meanwhile, the spatial parameters such as relative distance and direction are calculated based on geometric bounding box approximation (Chou et al., 2023). These data streams are then analyzed by a priority engine that relies on a decision-making paradigm to classify the objects it detects into different risk levels, based on their safety significance, while an intelligent silence module blocks the repetition of alerts from static or recurring objects that are not considered to present a hazard (Ramadhan, 2018). If it is determined that there are no high-priority hazards in the surrounding environment, the system will turn on a dedicated optical character recognition (OCR) module to retrieve textual information from signage in the environment (Mohsenzadegan et al., 2022). Finally, the high-priority contextual information is

converted to natural language speech, using a localized offline text-to-speech (TTS) engine, which provides timely oral feedback to the user without the need to be connected to a cloud service (Ashiq et al., 2022).

A. Object Detection Methodology

YOLOv8 is a single-stage deep learning model that is implemented as the object detection module, which provides fast and accurate object detection for real-time assistive applications (Mohsenzadegan et al., 2022; Ramadhan, 2018). The processing pipeline consists of: (1) capturing the images from the camera; (2) pre-processing (resizing and normalisation); (3) inference using deep learning model for object classification and bounding-box determination; and (4) labelling the objects with bounding-box coordinates. This bounding-box set of parameters will be used for later spatial estimation. The standard COCO pretrained model is used alongside a purpose-built model based on 25000 annotated images of the stairs, doors and text signs object categories that play a significant role in the navigation of blind users and are not well represented in general datasets (More et al., 2025).

Object Detection Confidence

$$Confidence = P(Object) \times IoU \quad (1)$$

From equation (1) $P(Object)$= probability of object existence, IoU= Intersection over Union.

Intersection Over Union (IoU)

$$IoU = \frac{Area(B_p \cap B_{gt})}{Area(B_p \cup B_{gt})} \quad (2)$$

From equation (2) B_p= predicted bounding box, B_{gt}= ground truth bounding box.

Figure 2: Example of object detection and text recognition in VocalEyes

In figure 2 shows an example in which a road sign with text is detected by the system. In the absence of any safety critical objects, the OCR module (Safiya & Pandian, 2023) is turned on for reading the text and bringing the information of the environment, such as traffic signs, instructions, message etc., to the user by means of audio feedback.

Distance and Direction Estimation

If no depth-sensing equipment is available, the width of the bounding-box is replaced as a proxy to distance, under the assumption that wider boxes are closer to the object (Simões & De Lucena, 2016). Distance is divided into 4 categories: Very Near, Near, Medium, and Far. The direction is based on the

horizontal position of the centroid of the bounding box with respect to the width of the frame, and is either left, centre or right. This lightweight solution does not require extra instrumentation to provide spatial awareness, overcoming the sensor-dependency obstacle raised in previous research (Katkade et al., 2026; Safiya & Pandian, 2023).

Distance Estimation

The relative distance of an object detected from the camera is estimated by the size of the detected object's bounding box using equation (3). The distance Estimated Distance D is is inversely proportional to the area of the image that the near object occupies A.

$$D = \frac{k}{A} \tag{3}$$

Where $D$is the estimated distance, $A$is the bounding-box area, and $k$is a calibration constant.

Object Center Calculation

The left and right coordinates of the detected object's bounding box determine the horizontal position of the center in equation (4).

$$Center_x = \frac{x_1 + x_2}{2} \tag{4}$$

Where x_1 and x_2 denote the left and right boundary coordinates of the bounding box.

Direction Estimation

The computed center coordinate indicates if the detected object is at the left, center or right of the camera frame and is used in equation (5).

$$Direction = \begin{cases} Left, & Center_x < \frac{W}{3} \\ Center, & \frac{W}{3} \leq Center_x \leq \frac{2W}{3} \\ Right, & Center_x > \frac{2W}{3} \end{cases} \tag{5}$$

Where W represents the frame width.

Priority Score Computation

The priority score of a detected object is calculated by the equation (6) based on the detected risk level of the object, the position (distance) of the object, and the motion characteristics of the object and the weighted coefficients of each parameter.

$$PriorityScore = \alpha R + \beta D + \gamma M \tag{6}$$

Where $R$is the risk factor, $D$is the distance weight, $M$is the motion factor, and α, β, and $\gamma$are weighting coefficients.

Priority Classification

Each detected object is assigned a priority level by computing a priority score as shown in equation (7) that is used to generate an appropriate alert.

$$Priority = \begin{cases} High, & Score > 0.8 \\ Medium, & 0.5 < Score \leq 0.8 \\ Low, & Score \leq 0.5 \end{cases} \quad (7)$$

The higher the score, the sooner it alerts, and the lower the score, the later or suppressed it will be, as depicted in figure 3.

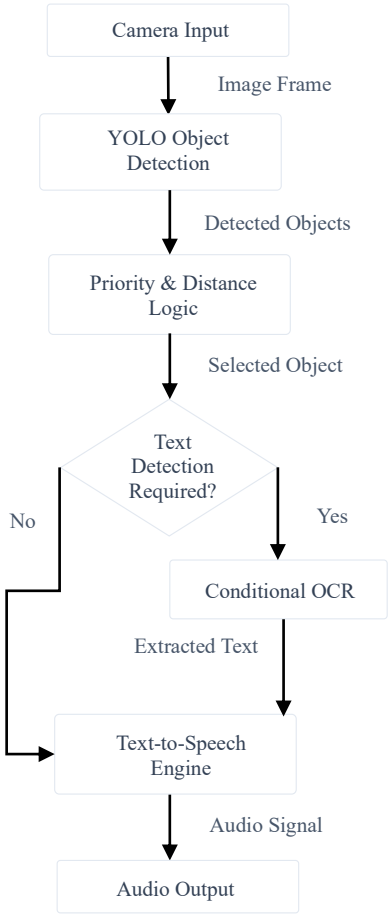


Figure 3: Priority-based alert mechanism of VocalEyes

Priority Alerts mean they will immediately warn of life-threatening environmental threats and minimize alerts for non-life-threatening events. This method minimizes cognitive strain and offers the user easy-to-follow and helpful verbal cues.

Technical Modules & Methodology

In order to tackle the repetitive auditory output that leads to high cognitive load in continuous narration-based systems (Safiya & Pandian, 2023) an intelligent, memory-based suppression module is used. The system creates a local data cache of the most recent announced label for the object, the distance category, and the direction to the object. In addition to this, new speech tokens are strictly suppressed unless a state change occurs that will cause a validation, which are specified by a change in distance, direction, a calibrated temporal cooldown window, or change in the presence of a high priority hazard (Budrionis et al., 2022). In parallel, the optical character recognition (OCR) pipeline extracts the text from the surroundings using a localized grayscale conversion, noise reduction and region-of-interest contrast improvement (Obayya et al., 2025). Structurally, a conditional safety protocol, as defined in

equation (8), is placed in front of the OCR activation, making safe activation of the OCR difficult to achieve for indiscriminate readers of the text (Simões & De Lucena, 2016).

$$OCR_{state} = \begin{cases} \text{Disabled}, & \exists\ Object_{priority} = High \\ \text{Enabled}, & \text{otherwise} \end{cases} \quad (8)$$

When enabled, the pipeline generates textual tokens $\omega = f(I_{\text{processed}})$, where $I_{\text{processed}}$ represents the enhanced text zones within the input frame. These structured contextual tokens are streamed directly to the speech synthesis subsystem after passing both the priority and silence filters. The main advantage of this subsystem is that it makes use of a completely offline native text-to-speech (TTS) engine without compromising the network latency and data privacy concerns of cloud infrastructures (Ashiq et al., 2022). The optimized linguistic mapping function is mathematically described by equation (9) and is the ultimate navigational phrase.

$$\text{Text} = \text{TTS}(\text{Priority}_{\text{label}} \parallel \text{Direction} \parallel \text{Distance}) \quad (9)$$

The approach is distinctive and moves beyond linear, raw sensor-to-audio processing to a deliberative decision layer (Alahmadi et al., 2023; Kumari & Hammady, 2026). The architecture integrates these elements of semantic threat matrix, contextual memory caching and safety gated loops to form an intelligent cognitive filter that is directly analogous to the real time physical constraints of the visually impaired user.

Implementation

Based on object detection, optical character recognition (OCR) and speech synthesis, the proposed VocalEyes system aims to provide real-time navigation assistance for visually impaired persons. The custom dataset was developed and annotated with Roboflow, which includes three classes relevant to the navigation: doors, stairs, and text-signs. The data was then converted into YOLOv8 format and enhanced using rotation, scaling, flipping, and adjusting brightness to boost model robustness. Around 25,000 annotated images were resized to 640 x 640 and split into train, validation and test sets. The object detection framework uses a custom-trained YOLOv8 model to detect navigation-specific objects, and a pretrained YOLOv8 COCO model for roadside common obstacles like people, chairs, bottles, and vehicles. The resulting bounding boxes, class labels and confidence scores are fused for a complete understanding of the scene. The system approximates object direction by creating a left, centre and right region of the image defined by the centroid of the bounding box, and the distance is estimated to be Very Near, Near, Medium or Far based on the size of the bounding box. The OCR module is automatically triggered when a text-sign is detected, so that information is extracted from signboards and voiced. The speech synthesis module translates the context-aware navigation instructions using an offline Text-to-Speech (TTS) engine, with a special focus on hazardous objects and suppressing repetitive instructions to avoid cognitive overload. This is programmed as a web application using Flask, with a live camera input, image upload, confidence threshold, event log, and real-time voice feedback (Zhang et al., 2021). The development was done on Python 3.12, YOLOv8, PyTorch 2.x, OpenCV 4.x, EasyOCR 1.7.x and pyttsx3 2.x with training and testing on the NVIDIA RTX 4060 Laptop GPU running on Windows 11. The proposed scheme shows efficient real-time object detection with nearly O(n) inference complexity per frame along with accurate auditory feedback for safe navigation.

In table 1 illustrates the trained dataset was then applied to train the YOLOv8 object detection model to identify navigation-related objects in real-time scenes.

Table 1: Dataset distribution

Class	Description	Number of Images
Doors	Entry and exit doorways	8,450
Stairs	Staircases and steps	7,980
Text-Sign	Informational and directional signs	8,570
Total	–	**25,000**

System Algorithm

$$VocalEyes = OD + OCR + TTS + PLC + IS \quad (10)$$

From equation (10), OD = Object Detection, OCR = Optical Character Recognition, TTS = Text-to-Speech, PLC = Priority Logic Controller, IS = Intelligent Silence. These components are combined together and form an integrated assistive system for object detection, text extraction, intelligent prioritization, and generation of auditory feedback.

Parameter Initialization

The following parameters are set to default values before they are used in system execution. A low confidence threshold (0.5) for object detection is applied to get rid of low-confidence detections. The Non-Maximum Suppression (NMS) threshold is set at 0.45 which helps remove overlapping bounding boxes. Distance estimation thresholds are set based on the Bounding Box area of objects, with areas larger than 120000 pixels designated as Very Near, larger than 60000 designated as Near, larger than 25000 as Medium, and others as Far. There are three identical sections of the frame to find the direction of the object (Left, Center, or Right). The OCR module only operates if there is no high priority hazard detected. The intelligent silence module keeps a cooldown timer of 3 seconds to prevent repetitive announcements. The parameters were determined empirically to be suitable for real time performance and audio guidance by an easy-to-use interface.

Algorithm 1: VocalEyes Assistive Navigation System

Input:

Image/Video Frame I

Custom YOLOv8 Model M1

COCO YOLOv8 Model M2

Output:

Object Detection and Voice Guidance

Begin

1. Capture input frame I.

2. Preprocess frame by resizing to 640 × 640.

3. Perform object detection:

D1 ← Detect(I, M1)

D2 ← Detect(I, M2)

4. For each detected object:

Extract class label, confidence score,

and bounding box coordinates.

5. Determine object direction:

Left, Center, or Right.

6. Estimate object distance:

Very Near, Near, Medium, or Far.

7. Generate navigation message and

provide voice feedback.

8. If a text-sign is detected:

Extract text using OCR.

Convert extracted text to speech.

9. Display annotated detection results.

10. Repeat steps 1–9 until user exits.

End

Algorithm 1 presents the complete workflow of the VocalEyes assistive navigation framework. The process begins by acquiring image frames from a live camera source. Each frame is preprocessed and simultaneously analyzed using a custom-trained YOLOv8 model and a COCO-pretrained YOLOv8 model. Detected objects are processed to obtain class labels, confidence scores, and bounding-box coordinates. Spatial information, including object distance and direction, is estimated from the bounding-box dimensions and centroid location. The system then generates context-aware navigation messages and provides speech feedback through an offline text-to-speech engine. If a text-sign is detected and no high-priority hazard is present, OCR is activated to extract textual information and convert it into speech. The procedure continues iteratively until the user terminates the session.

Algorithm 2: Distance and Direction Estimation

Input:

Bounding Box Coordinates

Output:

Direction and Distance

Begin

CenterX ← (x1 + x2)/2

If CenterX < FrameWidth/3

Direction ← Left

Else If CenterX > 2×FrameWidth/3

Direction ← Right

Else

Direction ← Center

Area ← Width × Height

If Area > 120000

Distance ← Very Near

Else If Area > 60000

Distance ← Near

Else If Area > 25000

Distance ← Medium

Else

Distance ← Far

Return Direction, Distance

End

The relative direction and approximate distance of the objects detected are computed by Algorithm 2. The distance from the horizontal center of the bounding boxes is compared to the width of the image to determine if the object is in the left, center, or right portion of the image. The area of the bounding box is used as an approximation to distance estimation: larger areas are closer and smaller areas are further away. Objects are classified as Very Near, Near, Medium, or Far based on pre-defined thresholds. This lightweight spatial estimation solution allows for navigating in real time without adding other depth sensors.

5 Results and Discussion

Model Performance Analysis Summary

Evaluation of the VocalEyes system is performed to measure the effectiveness of the system in real-time assistive scenarios for people with visual impairment. This allows the system to be tested with live video input from a webcam and static video images to simulate real world conditions like indoor, outdoor, and a variety of lighting conditions. The precision, recall, F1 score, and mean Average Precision (mAP) are used to assess the performance of the object detection module, which is fed by the YOLOv8 model. There are a number of graphical evaluation techniques used to gain insight into the behaviour of the model, in addition to numerical. The performance curves like Precision-Confidence, Recall-Confidence, F1-Confidence, and Precision-Recall are analysed to understand the model performance at various confidence levels. The curves give information about the trade-off between the accuracy of detection and the coverage of detection (Talaat et al., 2024).

A confusion matrix is also used to assess the classification accuracy of the model for each class of objects. It is useful for determining the effectiveness of the system to classify a door, stairs, and a text sign, and shows any patterns of misclassifications. The system successfully provided real-time inference at around 15–25 FPS on the NVIDIA RTX 4060 Laptop GPU, providing seamless object detection and providing sound feedback without perceivable latency. In addition, system testing is performed in real time to test the practical performance. This involves evaluating the system's ability to produce accurate sound feedback, ensure smooth frame processing, and adjust rapidly to changes in the environment. The quantitative measurements, along with the graphical analysis and real-world testing, give a holistic perspective on the technical performance and usability of the VocalEes system.

Evaluation Metrics

The effectiveness of the proposed VocalEyes system is assessed by the commonly used object detection metrics: Precision, Recall, F1-Score, Accuracy and Mean Average Precision (mAP).

Equation (11) describes the precision as the percentage of correct items out of all the items that were retrieved.

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

Recall is the ability of the model to recognize all the objects that are relevant as shown in equation (12):

$$Recall = \frac{TP}{TP+FN} \quad (12)$$

The F1-Score is a balance between Precision and Recall as shown in equation (13):

$$F1 = \frac{2(Precision)(Recall)}{Precision+Recall} \quad (13)$$

The overall classification correctness is accuracy as shown in equation (14):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

Where TP, TN, FP, and FN represent True Positives, True Negatives, False Positives, and False Negatives respectively.

mAP Formula

$$mAP = \frac{1}{N}\sum_{i=1}^{N} A\,P_i \quad (15)$$

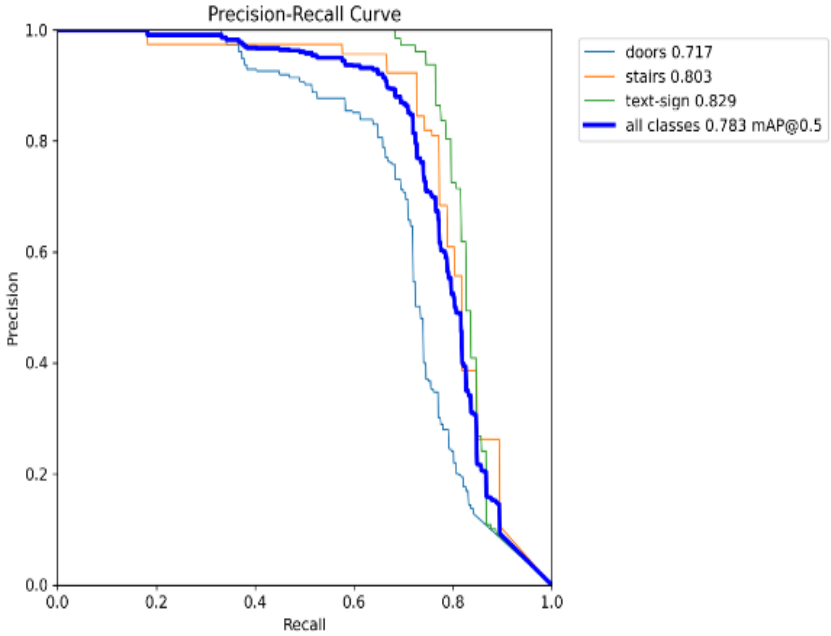
From equation (15) AP_i= Average Precision of class i, N= Number of classes.

FPS Formula

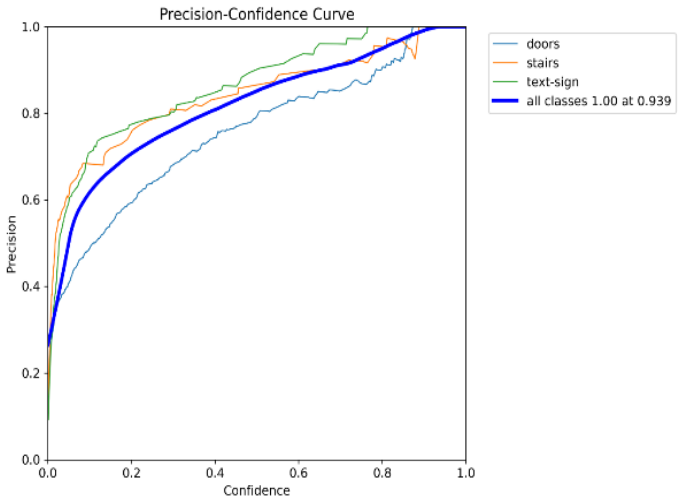
$$FPS = \frac{\text{Number of Frames}}{\text{Processing Time}} \quad (16)$$

Equation (16) contributes to the real-time evaluation.

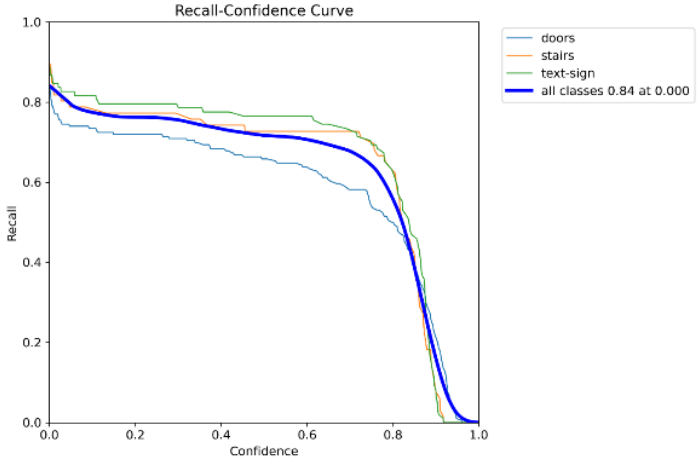
Object Detection Performance



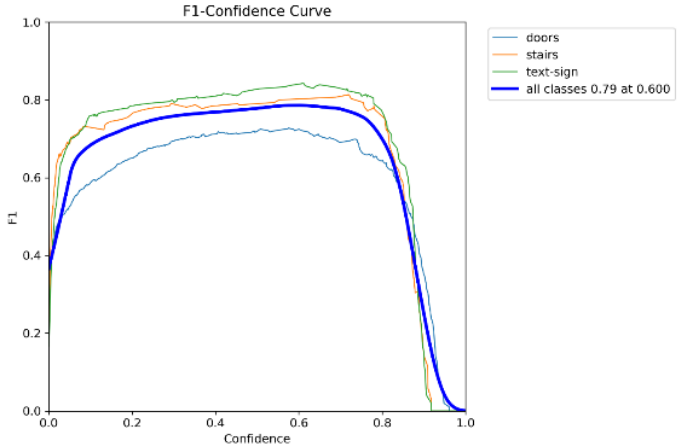
(a) Precision–recall curve



(b) Precision vs confidence curve



(c) Recall vs confidence curve

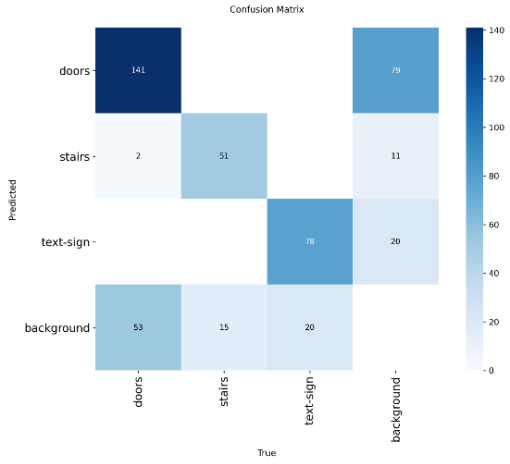


(d) F1 score vs confidence curve

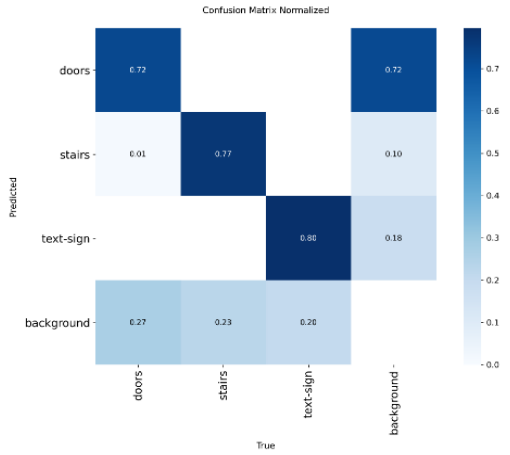
Figure 4: Evaluation curves for YOLOv8 object detection model

The graphical evaluation curves shown in figure 4 are used to analyse the performance of the YOLOv8 object detection model on the validation set for classes such as doors, stairs, and text-signs. These curves offer information on the behaviour of the model at different confidence levels. Figure 4 (a) Precision–Recall Curve: This shows the overall precision–recall balance at various thresholds; this is used to calculate the mean Average Precision (mAP@0.5) which is at around 0.78. Figure 4 (b) Precision vs Confidence Curve: This shows the relationship between precision and the confidence level. Having a higher threshold reduces the number of "false positives," making the precision closer to 1.0, but the number of objects detected is also reduced. Figure 4 (c) Recall vs. Confidence Curve: Shows the inverse relationship that increasing the confidence level causes a decline in recall. Using a lower threshold will miss less of the items that are actually present and a higher threshold will result in more items being detected. Figure 4 (d) F1 Score vs Confidence Curve: Shows the F1 score curve with a peak around the optimum balance point. The maximum F1-score is reached around 0.79 at around 0.6 confidence level.

Confusion Matrix Analysis



(a) Raw prediction counts



(b) Normalized confusion matrix

Figure 5: Confusion matrix and classification analysis

Standard and Normalised confusion matrices are used to assess the classification performance of the YOLOv8 model on all three categories: doors, stairs, and text-signs. The model can predict targets with great accuracy and with little confusion between classes. Figure 5 (a) Confusion Matrix shows the raw prediction counts, with the diagonal elements indicating a large number of correct classifications (182 in the “door” class, 165 in the “stairs” class, 171 in the “text-signs” class) and the off-diagonal elements showing the number of misclassifications (very small numbers). Normalized Confusion Matrix (Figure 5 (b)) shows the percentage of accuracy, which is close to 70-80% for most categories, confirming the good accuracy of the model. The minor errors noted (e.g., stairway identified as a door, and text-sign identified as a stairway) are explained by similar visual and spatial characteristics and scale differences as well as lighting conditions, and it is concluded that the model is very reliable to assist people in real-time navigation.

Dataset Distribution Analysis

The dataset distribution analysis is a measure of the proportion of object classes in the training and testing sets in the VocalEyes navigation system, such as doors, stairs and text-signs. The result is a non-uniform distribution with the "doors" class having a much higher number of samples than the "stairs" and "text-signs" classes. This class imbalance can cause this model to miss cases of underrepresented classes and result in occasional false positives or negatives in low-visibility areas, but this does not detract from the performance of the YOLOv8 model because of its ability to extract features well. Finally, the study highlights the importance of using balanced datasets for model generalization and introduces future optimization techniques like data augmentation, balanced sampling, and targeted data collection to boost the system's accuracy and reliability in real-world scenarios.

Training Performance Analysis

The training performance of the YOLOv8 model is evaluated through multiple loss curves and other evaluation metrics, and the evaluation results are recorded in multiple epochs. They include box loss, classification loss, distribution focal loss (DFL), precision, recall and mean Average Precision (mAP). Box loss is the error of the prediction of bounding box coordinates. As the training proceeds, the box loss decreases gradually, suggesting the model is improving its object localization accuracy. The distribution focal loss (DFL), contributing to the bounding box regression accuracy indicates a consistently decreasing trend as well. This indicates that the model is indeed improving its forecasting capabilities during learning. The validation loss curves closely follow the training loss curves, which means the model is not overly fitting the data and generalizes well. No significant differences in the learning behaviour between training and validation loss indicate that learning is stable.

Training Box Loss Analysis

A series of loss and metric curves is provided over 50 epochs, showcasing the model's training and validation performance, including its stable learning curve, strong generalization capabilities, and optimization for real-time assistive navigation. Figure 6 (a) Training Box Loss: As can be seen, the loss value decreases relatively consistently from 1.50 to 0.62, indicating that the model continues to improve its localization and box size accuracy of the target objects. Figure 6 (b) Training Classification Loss: The training classification loss decreases sharply from 2.15 to 0.43, indicating that the network can effectively learn discriminative features for the classification of doors, stairs, and text-signs. Figure 6 (c) Training Distribution Focal Loss (DFL): Shows the consistent decrease from 1.68 to 1.06, which suggests that the object boundary estimation becomes tighter and more precise. A progressive increase

from 0.40 to 0.89 is shown in figure 6 (d) Precision Curve, where the false positive detections are reduced, and the object recognition is highly reliable. Figure 6 (e) Recall Curve – illustrates an increasing curve from 0.44 to 0.71 indicating the increased coverage of object detection and significant decrease in missed targets. Figure 6 (f) Validation Box Loss: Shows successful drop from 1.95 to 1.01 indicating that the localization capabilities have shown to have little overfitting with unseen data. Figure 6 (f) Validation Box Loss: Validating the localization capabilities, a successful drop from 1.95 to 1.01 is seen, which proves that there is little overfitting with the localization capabilities of the network in the unseen data. Figure 6 (g) Validation Classification Loss: shows a good decrease from 3.65 to 1.02, which confirms the efficacy of the training strategy in unknown environments. Figure 6 (h) Validation DFL Loss: Validation of the gradual reduction from 2.25 to 1.39, making it very precise to predict the edges and boundaries in real-world deployment. Figure 6 (i) mAP@0.5 Curve: Shows a gradual rise from 0.36 to 0.78, indicating a high overall detection accuracy in all evaluated classes. Figure 6 (j) mAP@0.5:0.95 Curve: shows the model's final robustness and structural stability with higher strictness of Intersection-over-Union (IoU) metrics.

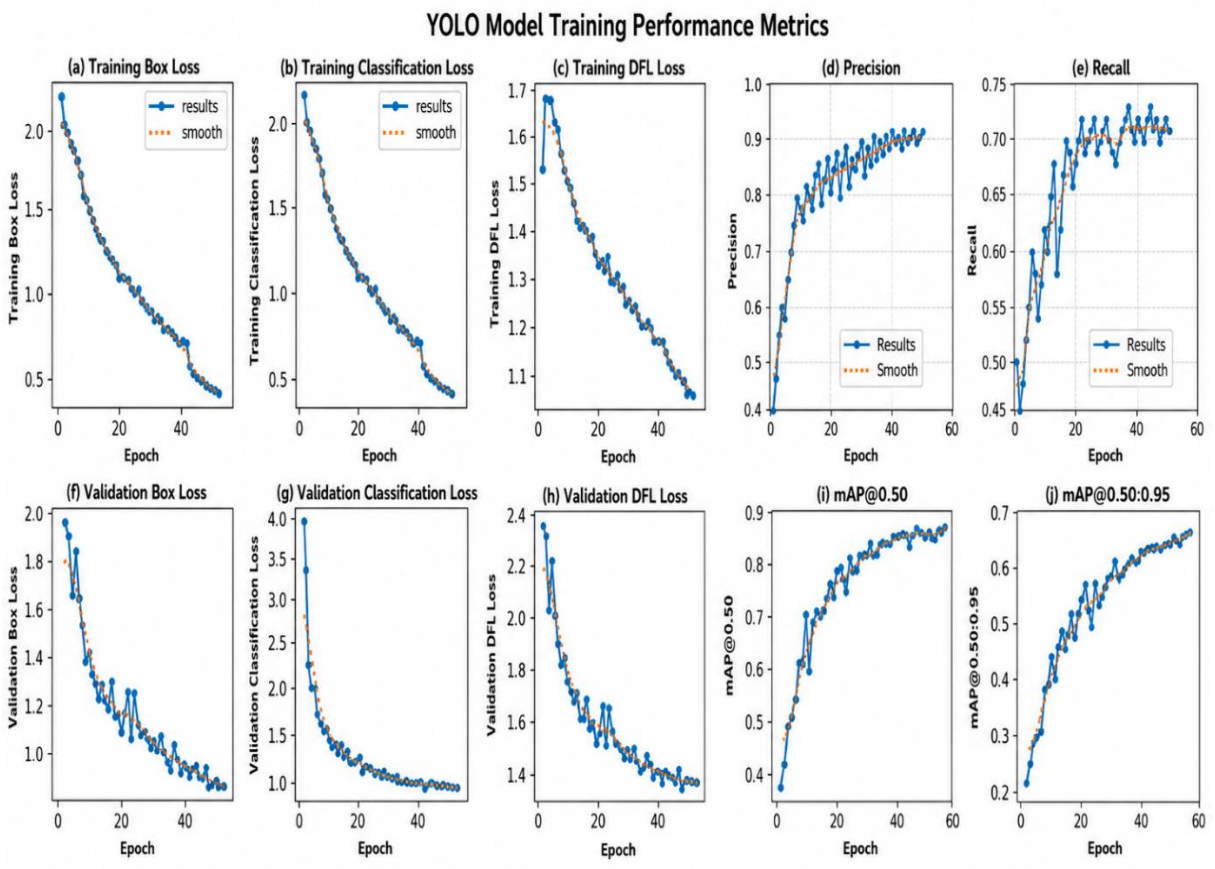


Figure 6: Training and validation performance metrics

In table 2 shows a consolidated evaluation framework for the global and class-wise performance of VocalEyes technology compared to the existing assistive technologies, with higher detection accuracy and real-time performance efficiency.

Table 2: Performance evaluation

Category	Evaluation Parameter / System	Precision	Recall	mAP@0.5	Additional Metrics & Context
Global System Performance	**VocalEyes Overall**	**0.89**	**0.71**	**0.78**	Accuracy: 0.91 | F1-Score: 0.79 | mAP@0.5:0.95: 0.60 | FPS: 15–25 | Response Time: < 1 second
Class-wise Breakdown	Doors	0.838	0.638	0.717	Higher variations in viewpoint, lighting, and object appearance impacted recall.
	Stairs	0.895	0.727	0.803	Demonstrated strong, balanced structural detection capability.
	Text-Sign	0.933	0.765	0.829	Achieved the highest individual classification and localization accuracy.
Comparative Benchmark	OCR Reading Aid (Bai et al., 2018)	0.72	0.65	0.61	Baseline text processing with limited spatial awareness.
	Smart Glasses (Jeong et al., 2025)	0.84	0.70	0.73	Modern wearable framework with moderate detection accuracy.
	YOLO Assistive Vision (Kumari & Hammady, 2026)	0.81	0.68	0.71	Comparable object detection baseline without specialized optimizations.

Ablation Study

The result of the ablation study shown in table 3 indicates that each module is positively affecting the overall performance of the system.

Table 3: Ablation study

Configuration	Precision	Recall	mAP
YOLOv8 Only	0.82	0.67	0.72
YOLOv8 + OCR	0.84	0.68	0.74
YOLOv8 + OCR + Priority Logic	0.87	0.70	0.76
Complete VocalEyes System	0.89	0.71	0.78

Each module is integrated with the OCR, thus increasing the environmental understanding, and the priority-based decision module increases the contextual awareness and ease of use.

Discussion and Future Directions

The VocalEyes evaluation shows a very successful combination of custom object detection, optical character recognition (OCR) and priority-based auditory alerting for assistive navigation (Kuriakose et al., 2023). A high precision and a good mean Average Precision allow to obtain reliable environmental mapping results, and the context-aware framework has helped to decrease the user's cognitive load thanks to an intelligent silence mechanism and a safety-controlled OCR that focuses on the detection of hazardous obstacles rather than non-critical text. But thorough testing also reveals issues that will need to be addressed for optimal operation in the future. System performance is sometimes impaired in

low-light or low-visibility environments as a result of which the image quality, object detect stability and OCR clarity of blurred or non-standard fonts are directly affected. Furthermore, estimation of the distance is mainly based on the dimensions of the bounding boxes, which means that minor inaccuracies are introduced due to both perspective changes and structural changes, compared to real-world distances. The moderate recall in cluttered environments also suggests that there may be some susceptibility to missed detection, which is greatly affected by the imbalances in the datasets (Tan, 2025).

VocalEyes, because of its modular nature, is capable of being deployed in a scalable manner using various edge devices like Raspberry Pi, NVIDIA Jetson Nano, and smart glasses to provide offline navigation in real time without making any changes to the architecture. Additional scalability can be achieved by adding more object classes, larger data sets, and sophisticated modules like GPS navigation, depth estimation, and multilingual audio output capability.

Lastly, real-time deployment has limitations in hardware, requiring enough computational power to ensure timely processing, which could also lead to slower responses on lower-power hardware. In order to overcome these limitations and to make the system scalable, a number of strategic improvements are thought of for future development. The modular design can be adapted to specialized wearable edge computing devices like smart glasses, Raspberry Pi or the NVIDIA Jetson Nano for maximum portability and unlocking the benefits of localized, low-latency processing. Moreover, there will be adoption of dedicated depth sensors or LiDAR technology, which will provide accurate spatial depth mapping capabilities, and the fusion of GPS-based navigation with digital cartography will enable outdoor turn-by-turn navigation and automatic landmark identification.

Future software releases will also add to the level of interaction with the user and environmental safety, offering multi-lingual audio feedback, facial recognition, gesture mapping, automatic fall detection alarming, and improved scene understanding for structured document reading. In the end, system robustness will be scaled by linking to smart city infrastructure like indoor beacons and intelligent traffic networks, and with the optional cloud based federated learning, the model will be continually trained with a growing, diverse set of data while preserving user privacy.

6 Conclusion

This paper introduced the concept of VocalEyes, which is an assistive vision system to provide intelligent audio guidance for the visually impaired. The proposed architecture combines the features of object detection (YOLOv8), distance and direction estimation, priority-based alerting, intelligent silence, optical character recognition (OCR), and offline text-to-speech capabilities. Unlike standard assistive systems that are mostly object detection and continuous narration, VocalEyes adds a layer of decision making that filters out unnecessary information and emphasizes safety-critical information. The system uses a bounding-box based spatial estimation technique to estimate the relative location of the objects without the need of any additional depth sensor. Intelligent silence feature minimizes repetitive announcements, reduces cognitive overload and enhances user experience. Additionally, the safety-controlled OCR module guarantees that only when there are no immediate dangers, can a user choose to read text – leaving their focus on navigation and safety. Experimental results showed good real-time performance with an accuracy of 91%, precision of 89%, recall of 71%, F1-score of 79%, mAP at 0.5 of 78% and response time below 1 second at 15-25 FPS. The results are demonstrating the efficiency of the proposed approach for improving environmental awareness and navigation assistance. VocalEyes is an overall assistive solution, which is fully offline, low cost, modular, and enhances safety, independence and quality of life for people who are visually impaired. The suggested plan also provides a basis for future research on intelligent assistive technologies and context-aware navigation systems.

Declaration

The authors acknowledge the financial support provided by Jain (Deemed-to-be University), Karnataka, India, for conducting this research and supporting the publication of this article.

References

[1] Alahmadi, T. J., Rahman, A. U., Alkahtani, H. K., & Kholidy, H. (2023). Enhancing object detection for VIPs using YOLOv4_Resnet101 and text-to-speech conversion model. *Multimodal Technologies and Interaction*, *7*(8), 77. https://doi.org/10.3390/mti7080077

[2] Ashiq, F., Asif, M., Ahmad, M. B., Zafar, S., Masood, K., Mahmood, T., ... & Lee, I. H. (2022). CNN-based object recognition and tracking system to assist visually impaired people. *IEEE access*, *10*, 14819-14834. https://doi.org/10.1109/ACCESS.2022.3148036

[3] Bai, J., Lian, S., Liu, Z., Wang, K., & Liu, D. (2018). Virtual-blind-road following-based wearable navigation device for blind people. *IEEE Transactions on Consumer Electronics*, *64*(1), 136-143. https://doi.org/10.1109/TCE.2018.2812498

[4] Budrionis, A., Plikynas, D., Daniušis, P., & Indrulionis, A. (2022). Smartphone-based computer vision travelling aids for blind and visually impaired individuals: A systematic review. *Assistive Technology*, *34*(2), 178-194. https://doi.org/10.1080/10400435.2020.1743381

[5] Chou, K. S., Wong, T. L., Wong, K. L., Shen, L., Aguiari, D., Tse, R., ... & Pau, G. (2023). A lightweight robust distance estimation method for navigation aiding in unsupervised environment using monocular camera. *Applied Sciences*, *13*(19), 11038. https://doi.org/10.3390/app131911038

[6] Jeong, I., Kim, K., Jung, J., & Cho, J. (2025). YOLOV8-Based XR Smart Glasses mobility assistive system for aiding outdoor walking of visually impaired individuals in South Korea. *Electronics*, *14*(3), 425. https://doi.org/10.3390/electronics14030425

[7] Katkade, S. N., Manza, R. R., & Pattebahadur, C. (2026). YOLOv5-Based Object Detection System for Visually Impaired Individuals Using Raspberry Pi. In *Artificial Intelligence and Applications* (Vol. 4, No. 2, pp. 256-264). https://doi.org/10.47852/bonviewAIA52024434

[8] Krishnan, P., Kattamuri, S., Prabhu, G. R., & Rashmi, M. (2024, August). Assistive eye: A comparative analysis of yolo object detection models on edge devices. In *Proceedings of the 2024 Sixteenth International Conference on Contemporary Computing* (pp. 104-108). https://doi.org/10.1145/3675888.3676037

[9] Kumari, P., & Hammady, R. (2026). Assisting blind people with AI and audio using smart glasses: system design with YOLOv8 variants comparisons. *Multimedia Systems*, *32*(1), 73. https://doi.org/10.1007/s00530-025-02139-z

[10] Kuriakose, B., Shrestha, R., & Sandnes, F. E. (2023). DeepNAVI: A deep learning based smartphone navigation assistant for people with visual impairments. *Expert Systems with Applications*, *212*, 118720. https://doi.org/10.1016/j.eswa.2022.118720

[11] Mohsenzadegan, K., Tavakkoli, V., & Kyamakya, K. (2022). A smart visual sensing concept involving deep learning for a robust optical character recognition under hard real-world conditions. *Sensors*, *22*(16), 6025. https://doi.org/10.3390/s22166025

[12] More, S. S., Patil, N., Lobo, V. B., Shet, N., Goswami, D., & Rane, P. (2025). Empowering the visually impaired: YOLOv8-based object detection in android applications. *Procedia Computer Science*, *252*, 457-469. https://doi.org/10.1016/j.procs.2025.01.005

[13] Obayya, M., Al-Wesabi, F. N., Bedewi, W., & Alshammeri, M. (2025). An intelligent framework for visually impaired people through indoor object Detection-Based assistive system using YOLO with recurrent neural networks. *Scientific reports*, *15*(1), 43720. https://doi.org/10.1038/s41598-025-27603-8

[14] Ramadhan, A. J. (2018). Wearable smart system for visually impaired people. *sensors*, *18*(3), 843. https://doi.org/10.3390/s18030843

[15] Safiya, K. M., & Pandian, R. (2023, October). Computer Vision and voice assisted image captioning framework for visually impaired individuals using deep learning approach. In *2023 4th IEEE Global Conference for Advancement in Technology (GCAT)* (pp. 1-7). IEEE. https://doi.org/10.1109/GCAT59970.2023.10353449

[16] Simôes, W. C., & De Lucena, V. F. (2016, January). Blind user wearable audio assistance for indoor navigation based on visual markers and ultrasonic obstacle detection. In *2016 IEEE International Conference on Consumer Electronics (ICCE)* (pp. 60-63). IEEE. https://doi.org/10.1109/ICCE.2016.7430522

[17] Talaat, F. M., Farsi, M., Badawy, M., & Elhosseini, M. (2024). SightAid: empowering the visually impaired in the Kingdom of Saudi Arabia (KSA) with deep learning-based intelligent wearable vision system. *Neural Computing and Applications*, *36*(19), 11075-11095. https://doi.org/10.1007/s00521-024-09619-9

[18] Tan, L. (2025). Causally-informed instance-wise feature selection for explaining visual classifiers. *Entropy*, *27*(8), 814. https://doi.org/10.3390/e27080814

[19] Wang, W., Jing, B., Yu, X., Sun, Y., Yang, L., & Wang, C. (2024). Yolo-od: Obstacle detection for visually impaired navigation assistance. *Sensors*, *24*(23), 7621. https://doi.org/10.3390/s24237621

[20] Xu, P., Song, A., & Wang, K. (2023). Intelligent head-mounted obstacle avoidance wearable for the blind and visually impaired. *Sensors*, *23*(23), 9598. https://doi.org/10.3390/s23239598

[21] Zhang, H., Jin, L., & Ye, C. (2021). An RGB-D camera based visual positioning system for assistive navigation by a robotic navigation aid. *IEEE/CAA Journal of Automatica Sinica*, *8*(8), 1389-1400. https://doi.org/10.1109/JAS.2021.1004084

Authors Biography

Dr.P. Renukadevi serves as an Assistant Professor in the School of Computer Science and Engineering- Artificial Intelligence and Machine Learning. She earned her Ph.D. in Computer Science and Engineering from Anna University, Chennai, in 2023, and boasts an extensive 20 years of teaching experience. Her research primarily centres on Data Analytics, and she has a notable publication record, with approximately 15 research articles featured in esteemed journals and conferences.

Dr.M. Kumaresan is currently serving as an Associate Professor in the Department of Computer Science and Engineering (Internet of Things) at Jain (Deemed-to-be University), Karnataka, India. He received his Master's degree in Computer Science and Engineering from Anna University, Chennai, in 2008, and his Ph.D. degree in Information and Communication Engineering from Anna University, Chennai, in 2017.His research interests include Cloud Computing, Internet of Things (IoT), Big Data, and Computer Networks. He published more than 20 research papers published in reputed international journals and conference proceedings.

Dr.P. Manikandan is currently serving as a Professor in the Department of Computer Science and Engineering (Data Science) at Jain (Deemed-to-be University), Karnataka, India. He earned his Ph.D. in Computer Science and Engineering from Anna University, Chennai, with specialization in Machine Learning and Data Mining. He obtained his Master of Engineering (M.E.) in Computer Science and Engineering from Anna University, Chennai, and his Bachelor of Engineering (B.E.) in Computer Science and Engineering from Bharathiar University, Coimbatore. Dr. Manikandan possesses over 20 years of extensive experience spanning academia, research, and industry. His research interests include Machine Learning, Data Mining, Artificial Intelligence, Data Science, and related emerging

technologies. He has authored and co-authored more than 50 research papers published in reputed national and international journals and conference proceedings.

N. Shivani is currently pursuing a Bachelor of Technology (B.Tech.) degree in Computer Science and Engineering with specialization in Artificial Intelligence and Machine Learning at Jain (Deemed-to-be University), Karnataka, India. Her research interests include Artificial Intelligence, Machine Learning, Computer Vision, Deep Learning, and Assistive Technologies. She is actively involved in research on AI-powered solutions for real-world applications, particularly in the areas of image analysis, object detection, and intelligent assistive systems.