International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 04 April-2026 | Impact Factor: 3.5

A Hybrid YOLO-Based Assistive Vision System for Visually Impaired users: Real-Time Object Detection and Voice Feedback

Ranganatha Sarma[1] and Dr. Abuzar Ansari[2]

Ranganatha Sarma, student of Department of Data Science, SIES College of Arts, Science and Commerce, Mumbai, India

How to Cite this Article:

Sarma, R. (2026). A Hybrid YOLO-Based Assistive Vision System for Visually Impaired users: Real-Time Object Detection and Voice Feedback. International Journal of Creative and Open Research in Engineering and Management, 02(04). https://doi.org/10.55041/ijcope.v2i4.444

https://doi.org/10.55041/ijcope.v2i4.444

Abstract — Visual impairment affects over 2.2 billion people worldwide, severely restricting independent living, particularly in indoor environments where object recognition and navigation remain major challenges. Traditional aids such as white canes offer limited semantic information, while commercial solutions like Microsoft Seeing AI and OrCam MyEye are often costly and may require internet connectivity. Recent advances in deep learning, especially single-stage object detectors like YOLO, have enabled real-time, on-device assistive technologies. This paper presents a hybrid assistive vision system that integrates a pretrained YOLOv11n detection model (for persons and general objects) with a custom fine-tuned YOLOv11n-cls classification model focused on five common indoor furniture items prevalent in Indian households: chair,television, refrigerator, table, and almirah (wardrobe). A dataset of 15,000 diverse images was curated under varied lighting, angles, and indoor conditions. The custom model achieved 99.61% accuracy,precision, recall, and F1-score on a held-out test set of 2,045 images.

The complete pipeline processes live webcam input using OpenCV and Ultralytics YOLO, delivering real-time performance of 12–18 frames per second (FPS) on a standard laptop CPU without requiring GPU hardware. Detected objects are conveyed via offline text-to-speech (pyttsx3) with a 2- second cooldown mechanism to avoid auditory overload. A preliminary qualitative user study with five participants reported an average satisfaction rating of 4.5/5, highlighting the system's voice clarity and practical utility for indoor navigation.

This work contributes a low-cost, customizable, fully offline solution tailored to Indian household environments. The modular hybrid architecture balances broad generalization with domain-specific precision, addressing key gaps in existing assistive technologies.

Keywords — Assistive technology, object detection, YOLO, visually impaired, indoor navigation, voice feedback, computer vision for accessibility, hybrid model

Ⅰ. Introduction

Visual impairment remains one of the most pressing global public health issues, impacting the daily independence of more than 2.2 billion individuals. Indoor navigation and real-time object recognition pose persistent difficulties, often forcing reliance on caregivers or basic tactile aids. While white canes and guide dogs provide essential mobility support, they deliver minimal contextual awareness about the surrounding environment, such as identifying furniture, appliances, or approaching people.

Commercial assistive devices have made progress, yet many remain expensive,dependent on cloud connectivity, or limited in scope. Recent breakthroughs in computer vision, particularly the YOLO (You Only Look Once) family of models, have demonstrated strong potential for lightweight, real-time object detection suitable for edge

deployment. However, most existing systems target outdoor navigation or generic object detection and lack customization for everyday indoor items commonly found in Indian households, such as the almirah (a traditional wardrobe).

This study addresses these limitations by proposing a **hybrid YOLO-based assistive vision system**. The system combines a pretrained YOLOv11n detection model for general objects and persons with a custom fine-tuned YOLOv11n-cls model specialized in five key furniture categories. By processing live video from a standard webcam and providing concise voice feedback, the system enhances situational awareness while remaining fully offline and cost-effective. The primary goal is to empower visually impaired users with greater independence in familiar indoor settings through accurate, timely, and non-intrusive auditory cues.

Ⅱ. Literature Review

The development of AI-assisted vision

systems for the visually impaired has grown rapidly, intersecting real-time object detection, indoor navigation, and voice-based feedback mechanisms.

Several studies have leveraged YOLO variants for assistive navigation.
Davanthapuram et al. proposed a YOLO- based indoor navigation framework that integrates object recognition with monocular depth estimation and binaural spatial audio, enabling more intuitive
guidance for users [1]. Wang et al.

introduced YOLO-OD, an enhanced model incorporating feature weighting and attention mechanisms to improve obstacle detection accuracy in navigation assistance scenarios [2]. Recent works have also explored YOLOv8 and YOLOv11 on embedded platforms, confirming their suitability for real-time performance under resource constraints [3], [4].

Domain-specific fine-tuning has proven effective for household object recognition. Noor et al. demonstrated the feasibility of YOLOv11 for real-time indoor object detection on Raspberry Pi 4, achieving high accuracy on furniture-related categories while maintaining acceptable inference latency [5]. Other approaches have combined YOLO with recurrent neural networks for sequential scene
understanding, improving stability across video frames in dynamic indoor environments [6].

Voice feedback systems emphasize the importance of clear, non-overwhelming auditory output. Temporal smoothing and cooldown mechanisms are frequently
recommended to reduce cognitive load.

Offline capability and low-cost deployment remain critical factors for widespread
accessibility, especially in developing regions [7].

Despite these advancements, significant gaps persist. Many systems lack
customization for region-specific furniture such as the almirah, rely on expensive hardware or internet connectivity, or fail to incorporate robust voice feedback logic.

This study bridges these gaps through a hybrid architecture that combines general detection with targeted classification, optimized for Indian indoor contexts, while ensuring fully offline operation on standard hardware.

III. Proposed Methodology

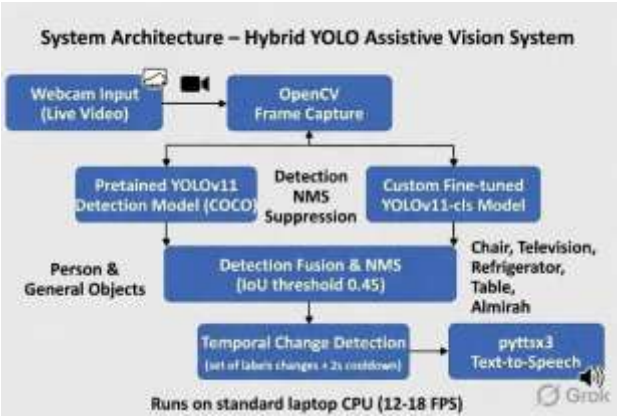
The proposed system employs a hybrid AI pipeline designed for real-time indoor assistance. It processes live video frames from a standard webcam and fuses outputs from two complementary YOLOv11 models.

A. System Architecture

The architecture consists of parallel processing branches. The pretrained YOLOv11n (trained on COCO) handles detection of persons and general everyday objects, providing broad environmental context. Simultaneously, the custom fine-tuned YOLOv11n-cls model classifies five specific furniture categories: chair, television, refrigerator, table, and almirah. Outputs are fused using Non-Maximum Suppression (NMS) with an IoU threshold of 0.45 to eliminate redundant detections.

A temporal change detection module monitors the set of detected labels across frames. Voice announcements are generated only when the detected label set changes and a 2-second cooldown period has elapsed. This design prevents repetitive or overlapping speech, thereby minimizing user fatigue.

Fig. 1. System Architecture of the Hybrid YOLO-Based Assistive Vision System. (The diagram illustrates the complete pipeline: webcam input captured via OpenCV, parallel branches for pretrained detection and custom classification, detection fusion with NMS, temporal change detection with cooldown, and pyttsx3 text-to-speech output. The entire pipeline runs efficiently on a standard laptop CPU at 12–18 FPS.)



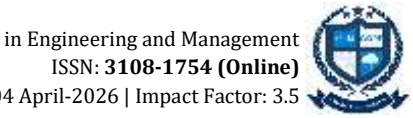
B. Dataset Acquisition and Preparation

A custom dataset of 15,000 labeled images was collected, with 3,000 images per furniture class. Images were captured under diverse indoor conditions typical of Indian households, including varying lighting (natural, artificial, low-light), viewing angles, partial occlusions, and background clutter. The dataset was divided into 80% training, 10% validation, and 10% test sets (2,045 test images).

Data augmentation techniques such as random rotation, brightness/contrast adjustment, and horizontal flipping were applied during training to enhance robustness.

C. Model Development and Training

• **Detection Model**: The pretrained YOLOv11n model (COCO weights) was used directly for detecting persons and general objects.

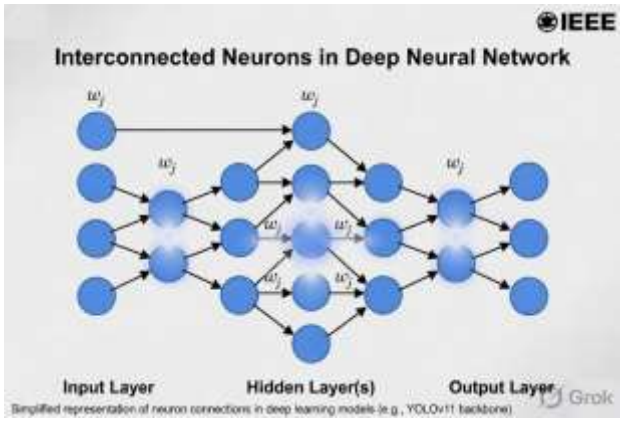


• **Classification Model**:

The YOLOv11n-cls variant was fine-tuned for 30 epochs on the custom dataset with image size 224×224 and batch size 8. Training was performed on a standard laptop with Adam optimizer.

D. Fusion, Post-Processing, and Voice Feedback

The system was developed in Python using the Ultralytics YOLO library, OpenCV for frame capture, and pyttsx3 for speech synthesis. No GPU acceleration was used during inference, achieving consistent 12–18 FPS on a standard laptop CPU. This **demonstrates the practicality of deploying advanced computer vision without specialized hardware.**



IV. Mathematical Modeling and Verification Logic

A. YOLO Output Decoding

YOLO employs anchor-based regression. For each anchor a, the normalized center coordinates and dimensions are decoded as:

$$c_{x_{norm}} = \frac{rawOutput[0 \times numAnchors + a]}{inputSize}, c_{y_{norm}} = \frac{rawOutput[1 \times numAnchors + a]}{inputSize}$$

$$w_{norm} = \frac{rawOutput[2 \times numAnchors + a]}{inputSize}, h_{norm} = \frac{rawOutput[3 \times numAnchors + a]}{inputSize}$$

offline pyttsx3 library. Announcements are triggered conditionally based on label set changes and the cooldown timer, ensuring natural and non-intrusive interaction.

E. Implementation Details

$$x_1 = \max(0, c_{x_{norm}} - w_{norm}/2), y_1 = \max(0, c_{y_{norm}} - h_{norm}/2)$$

$$x_2 = \min(1, c_{x_{norm}} + w_{norm}/2), y_2 = \min(1, c_{y_{norm}} + h_{norm}/2)$$

The bounding box corners are then computed as:

$x_1 = \max(0, c_{x\text{norm}} - w_{\text{norm}}/2), y_1$

$= \max(0, c_{y\text{norm}} - h_{\text{norm}}/2)$

$x_2 = \min(1, c_{x\text{norm}} + w_{\text{norm}}/2), y_2$

$= \min(1, c_{y\text{norm}} + h_{\text{norm}}/2)$

B. Non-Maximum Suppression (NMS)

For any two bounding boxes $A$and B, the Intersection over Union (IoU) is calculated as:

$$\text{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Boxes with IoU exceeding 0.45 are suppressed, retaining only the highest- confidence prediction.

C. Temporal Change Detection

Voice output is triggered only if the current label set $L_t$differs from the previous set $L_{t-1}$and the time elapsed since the last announcement satisfies $t - t_{\text{last}} \geq 2000$ms.

Table I Performance Metrics on Test Set

Metric	Value
Accuracy	0.9961
Precision	0.9961
Recall	0.9961
F1-Score	0.9961

V. Results and Discussion

The custom classification model achieved near-perfect performance (99.61% across accuracy, precision, recall, and F1-score) on the held-out test set. This high accuracy reflects effective learning of distinctive indoor furniture patterns under controlled conditions. Real-time inference reached 12– 18 FPS on standard laptop CPU, confirming suitability for practical assistive use without dedicated hardware.

The preliminary qualitative user study involving five participants in real indoor settings yielded positive feedback, with an average satisfaction rating of 4.5 out of 5. Participants particularly appreciated the clarity of voice announcements and the system's ability to identify commonly used household items, which enhanced their confidence during navigation.

Compared to pretrained models alone, the hybrid approach significantly improved recognition of domain-specific items like almirah. The 2-second cooldown mechanism effectively reduced redundant announcements, lowering cognitive load.

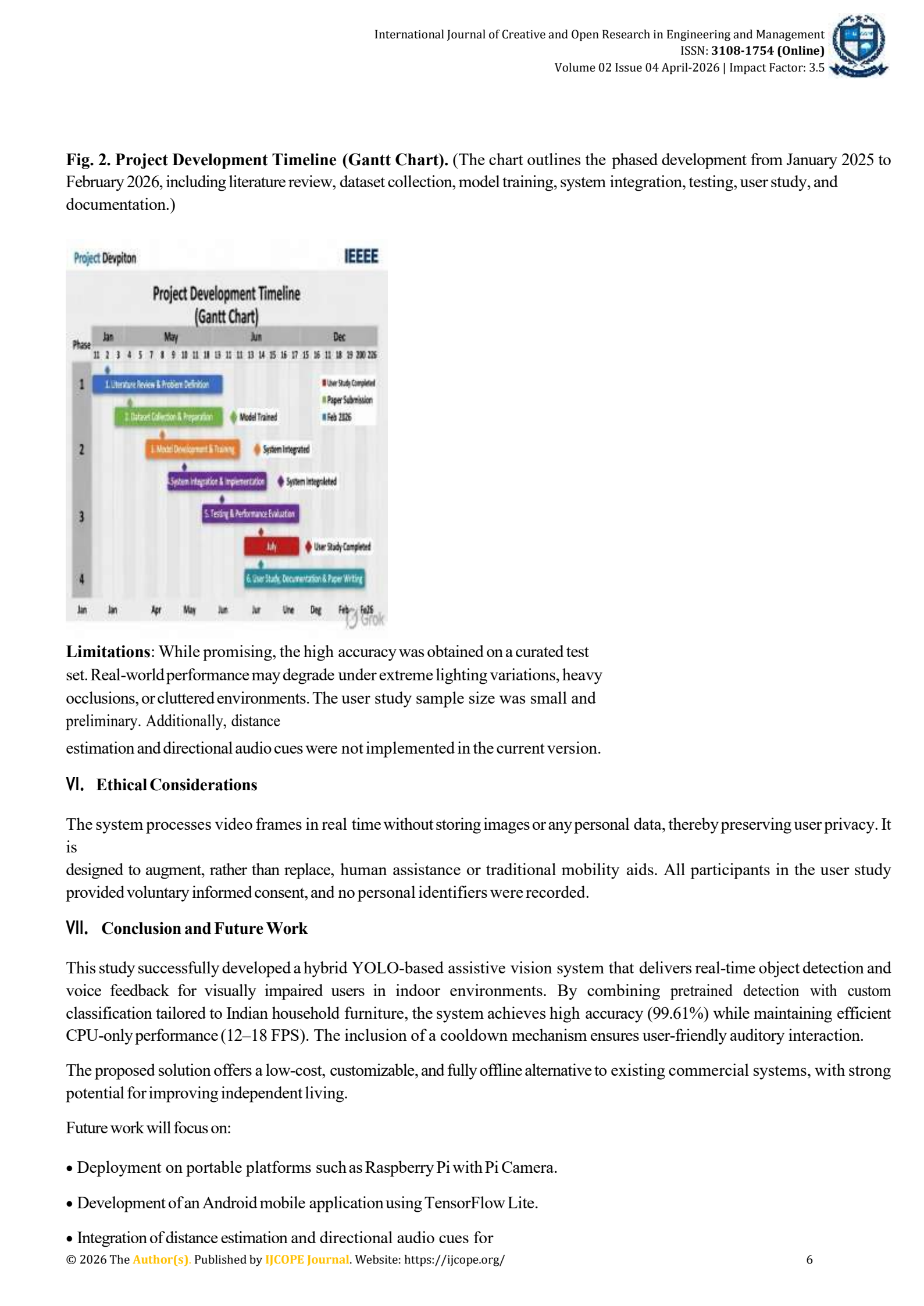


Fig. 2. Project Development Timeline (Gantt Chart). (The chart outlines the phased development from January 2025 to February 2026, including literature review, dataset collection, model training, system integration, testing, user study, and documentation.)

Limitations: While promising, the high accuracy was obtained on a curated test set. Real-world performance may degrade under extreme lighting variations, heavy occlusions, or cluttered environments. The user study sample size was small and preliminary. Additionally, distance estimation and directional audio cues were not implemented in the current version.

VI. Ethical Considerations

The system processes video frames in real time without storing images or any personal data, thereby preserving user privacy. It is designed to augment, rather than replace, human assistance or traditional mobility aids. All participants in the user study provided voluntary informed consent, and no personal identifiers were recorded.

VII. Conclusion and Future Work

This study successfully developed a hybrid YOLO-based assistive vision system that delivers real-time object detection and voice feedback for visually impaired users in indoor environments. By combining pretrained detection with custom classification tailored to Indian household furniture, the system achieves high accuracy (99.61%) while maintaining efficient CPU-only performance (12–18 FPS). The inclusion of a cooldown mechanism ensures user-friendly auditory interaction.

The proposed solution offers a low-cost, customizable, and fully offline alternative to existing commercial systems, with strong potential for improving independent living.

Future work will focus on:

- Deployment on portable platforms such as Raspberry Pi with Pi Camera.
- Development of an Android mobile application using TensorFlow Lite.
- Integration of distance estimation and directional audio cues for

enhanced spatial awareness.

- Training a unified custom detection model that natively includes the almirah class.
- Conducting larger-scale user studies with a diverse group of visually impaired participants to evaluate long-term usability and impact.

Overall, this work lays a strong foundation for accessible, region-specific indoor assistive technologies using modern computer vision techniques.

References

[1] S. Davanthapuram, X. Yu, and J. Saniie, "Visually Impaired Indoor Navigation using YOLO Based Object Recognition, Monocular Depth Estimation and Binaural Sounds," in Proc. IEEE Int. Conf. Electro Information Technology (EIT), 2021, pp. 173–177.

[2] W. Wang et al., "YOLO-OD: Obstacle Detection for Visually Impaired Navigation Assistance," Sensors, vol. 24, no. 23, Art. no. 7621, 2024, doi: 10.3390/s24237621.

[3] A. Noor et al., "Towards a Real-Time Indoor Object Detection for Visually Impaired Users Using Raspberry Pi 4 and YOLOv11: A Feasibility Study," Microprocess. Microsyst., 2025.

[4] M. Obayya et al., "An intelligent framework for visually impaired people through indoor object detection-based assistive system using YOLO with recurrent neural networks," Sci. Rep., vol. 15, Art. no. 43720, Dec. 2025, doi: 10.1038/s41598-025-27603-8.

[5] V. Hingnekar et al., "Netra AI: Real-Time AI-Powered Navigational Assistance forVisually Impaired Individuals UsingOptimized YOLOv11 Architecture," TechRxiv, 2025.

[6] X. Yu et al., "Visual Impairment Spatial Awareness System for Indoor Activities," J. Imaging, vol. 11, no. 1, Art. no. 9, 2025, doi: 10.3390/jimaging11010009.

[7] Microsoft, "Seeing AI – Talking Camera App for the Blind," 2025. [Online]. Available: https://www.microsoft.com/en-us/ai/seeing-ai