WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal Homepage: www.ijrpr.com ISSN 2582-7421

Spatialsense AI (Echomap): Persistent 3d Spatial Memory and AI Voice Navigation for Visually Impaired Users

Ms. Hemalatha[1], Divya[2], Baranisha[3], Akshaya[4], Bheeshma Sri[5]

Department of Artificial Intelligence and Data Science, Bachelor of Engineering, [College Name], Coimbatore, India-[PIN]

ABSTRACT :

SpatialSense AI (EchoMap) is an assistive software system that helps blind and visually impaired people understand their surroundings. Most existing vision applications identify objects in a single camera frame and then forget them. The proposed system detects objects with YOLOv8, tracks them across frames with ByteTrack/DeepSORT, places them in a lightweight room map using a mini-SLAM module, stores what it has seen as a persistent scene memory, and reports changes through a voice assistant that answers in the user's own language.

The system answers questions such as "What is in front of me?" and "Where is my bag?", detects objects that are new, moved or missing since the last scan, and clearly separates what is visible now from what was only seen earlier. Voice interaction is supported in Tamil, English, Telugu and Hindi. SpatialSense AI is intended as a supplement to, and not a replacement for, a white cane, guide dog or other established mobility aids.

Keywords: assistive technology, visual impairment, object detection, YOLOv8, object tracking, SLAM, scene memory, voice assistant, multilingual speech.

INTRODUCTION:

People with visual impairment often find it difficult to identify nearby objects, understand room layouts and notice changes in familiar places. A chair moved into a walking path or a bag left near a doorway can be a real hazard. Modern computer vision can recognise objects, but most applications describe a single image and carry no memory of the place.

The proposed SpatialSense AI (EchoMap) system combines four capabilities in one pipeline: real-time object detection, multi-frame tracking, spatial mapping and persistent scene memory. A voice interface delivers the result as short spoken sentences, so the user never needs to look at a screen.

The application processes a live camera stream, estimates the direction (left, centre, right) and approximate distance (near, medium, far) of each object, stores structured records for each room, and compares later scans with the stored scene. Responses are kept short, ordered by importance and honest about uncertainty.

OBJECTIVE:

- To detect common indoor objects (chair, table, door, bag, bottle, person, laptop, sofa and others) in real time.
- To estimate each object's direction and approximate distance from the camera view.
- To track objects across frames so that each one keeps a stable identity.
- To build a persistent spatial memory of a room and compare it with later scans.
- To detect added, removed and moved objects and announce them by voice.
- To provide a voice assistant that answers questions about the scene in Tamil, English, Telugu and Hindi.
- To never present stale information as a live detection.

LITERATURE SURVEY:

1. Redmon et al. (2016) introduced YOLO, which treats object detection as a single regression problem and processes a whole image in one pass. This makes real-time detection practical and is the basis of the YOLOv8 detector used in this work.

2. Bewley et al. (2016) proposed SORT, a simple online tracker that associates detections across frames using motion prediction. Wojke et al. (2017) extended it as DeepSORT by adding appearance features, which helps recover identities after short occlusions.

3. Zhang et al. (2022) presented ByteTrack, which associates both high- and low-confidence detection boxes instead of discarding the low-confidence ones. This improves tracking when objects are partly hidden.

4. Mur-Artal et al. (2015) and Cadena et al. (2016) described feature-based SLAM and reviewed the state of simultaneous localisation and mapping. Their work explains how camera motion can be estimated while a consistent map is built, which motivates the mini-SLAM module of this system.

5. Kuriakose et al. (2022) and Messaoudi et al. (2022) reviewed tools and technologies for navigation support for blind and visually impaired people. Both reviews show that many solutions focus on obstacle detection or single-frame description and give limited support for memory, change awareness and local-language interaction.

METHODOLOGY:

1. Frame Acquisition: On start-up the application requests camera permission and captures frames continuously without waiting for a button press. Frames are resized and passed to the detector at a controlled rate to balance responsiveness and processing load.

2. Object Detection: YOLOv8 returns a class label, a confidence score and a bounding box for each object. A confidence threshold removes weak detections, and only classes relevant to indoor navigation are announced.

3. Object Tracking: ByteTrack/DeepSORT links detections of the same object across frames and assigns persistent IDs. This prevents the same chair from being announced repeatedly and separates “the same bag moved” from “a second bag appeared”.

4. Spatial Mapping: The view is divided into left, centre and right regions for direction. Distance is expressed as near, medium or far from bounding-box position and size, because a single camera cannot give calibrated depth. The mini-SLAM module estimates camera motion and places tracked objects in a consistent room coordinate frame.

5. Scene Memory: Structured records (room identifier, object class, track ID, approximate map position, direction, distance band, timestamp and last confirmed detection) are stored instead of raw video. This keeps storage small and protects privacy.

6. Change Detection: On a new scan, the stored scene for that room is compared with current detections. Each object is classified as unchanged, moved, new or missing. Missing objects are reported cautiously, because they may only be outside the camera view.

7. Voice Interaction: Speech is converted to text, the Groq-hosted language model interprets the question against current detections and scene memory, and the answer is spoken through Text-to-Speech in the language the user spoke. Live and remembered information are always stated separately.

EXISTING SYSTEM:

1. Single-Frame Description: Most assistive vision applications describe one image at a time and do not link objects across frames.

2. No Memory of the Place: Existing tools do not store what was seen earlier, so they cannot answer “where was it before?”.

3. No Change Awareness: When a bag is moved from a table to a doorway, a basic detector recognises a bag in both cases but does not report that it moved.

4. Live and Old Data Mixed: Most tools make no distinction between currently visible and previously seen objects.

5. Limited Language Support: Many tools work mainly in English and give long descriptions that are hard to use while walking.

6. Connectivity Dependence: Many tools depend on a constant internet connection.

PROPOSED SYSTEM:

1. Continuous Multi-Frame Understanding: Detection and tracking run on a live stream with stable object IDs.

2. Direction and Distance: Each object is reported with left/centre/right direction and a near/medium/far distance band.

3. Persistent Scene Memory: Structured room records are stored and reused in later sessions.

4. Change Detection: Added, removed and moved objects are detected and announced.

5. Voice-First Interaction: The user asks questions by voice and hears short answers.

6. Multilingual Support: Tamil, English, Telugu and Hindi, with replies in the language spoken.

7. Honest Reporting: Live detections and remembered objects are always separated.

SYSTEM REQUIREMENTS:

Hardware Requirements:

- Processor: Intel Core i5 or higher, or equivalent. [Update to match your test device]
- RAM: Minimum 8 GB recommended. [Update]
- Camera: Built-in or external webcam / phone camera.
- Audio: Microphone and speaker or headphones.
- Network: Internet connection for the language-model API.

Software Requirements:

- Operating System: Windows 10/11, Linux or macOS.
- Frontend: React.js, HTML and CSS.
- Backend: FastAPI (Python).
- AI Processing: YOLOv8, ByteTrack/DeepSORT, mini-SLAM module.
- Language Model: Groq API.
- Speech: Speech-to-text and Text-to-Speech services.
- Browser: Google Chrome, Mozilla Firefox or Microsoft Edge.
- [Edit this list so it names only tools actually used in the final version.]

PROCESSING:

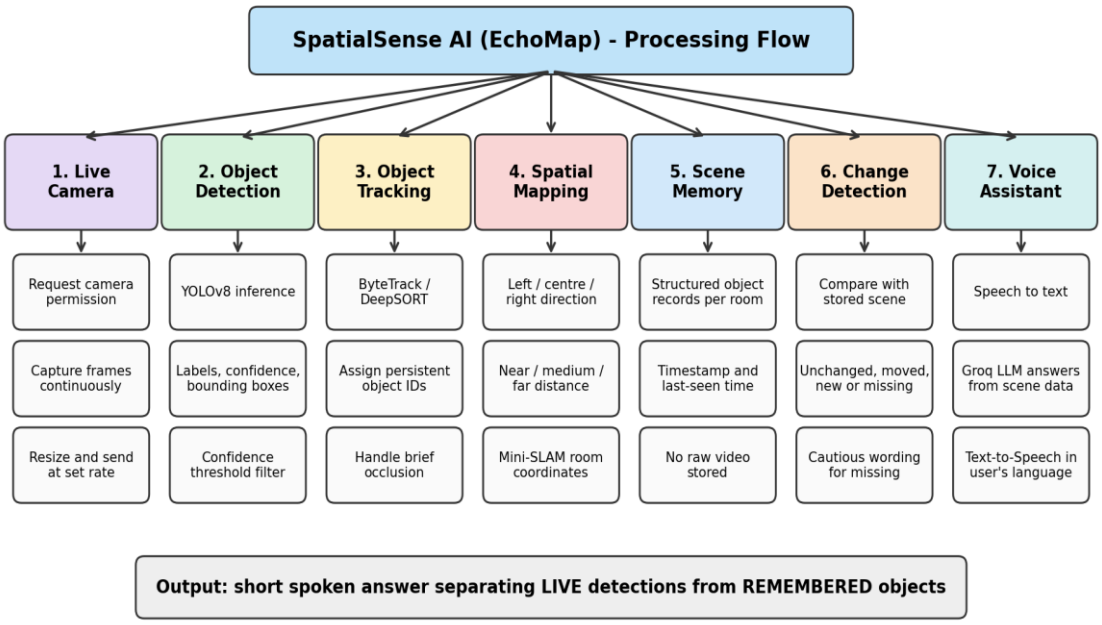


Fig. 1: Processing flow of SpatialSense AI (EchoMap)

MODULES:

1. Live Camera Module: Captures frames continuously and sends them to the detector at a controlled rate.

2. Object Detection Module: Uses YOLOv8 to return labels, confidence scores and bounding boxes.

3. Object Tracking Module: Uses ByteTrack/DeepSORT to maintain persistent object IDs.

4. Spatial Awareness Module: Computes direction, distance band and room coordinates using the mini-SLAM module.

5. Scene Memory Module: Stores and retrieves structured object records for each room.

6. Change Detection Module: Reports objects as unchanged, moved, new or missing.

7. Voice Assistant Module: Interprets questions with the Groq API and speaks answers using Text-to-Speech.

8. Multilingual Module: Supports Tamil, English, Telugu and Hindi input and output.

Change classification used by the system:

Outcome	Rule
Unchanged	Same class in approximately the same position.
Moved	Same object now in a different region or distance band.
New	Present now, not in memory.
Missing	In memory, not detected now (reported cautiously).

Example scene-memory record:

{ "room_id": "room_01",

"objects": [

{ "name": "chair", "direction": "left", "distance": "near", "last_seen": "2026-11-01T10:15:00" },

{ "name": "table", "direction": "front", "distance": "medium", "last_seen": "2026-11-01T10:15:00" }] }

APPLICATION SCREENSHOTS:

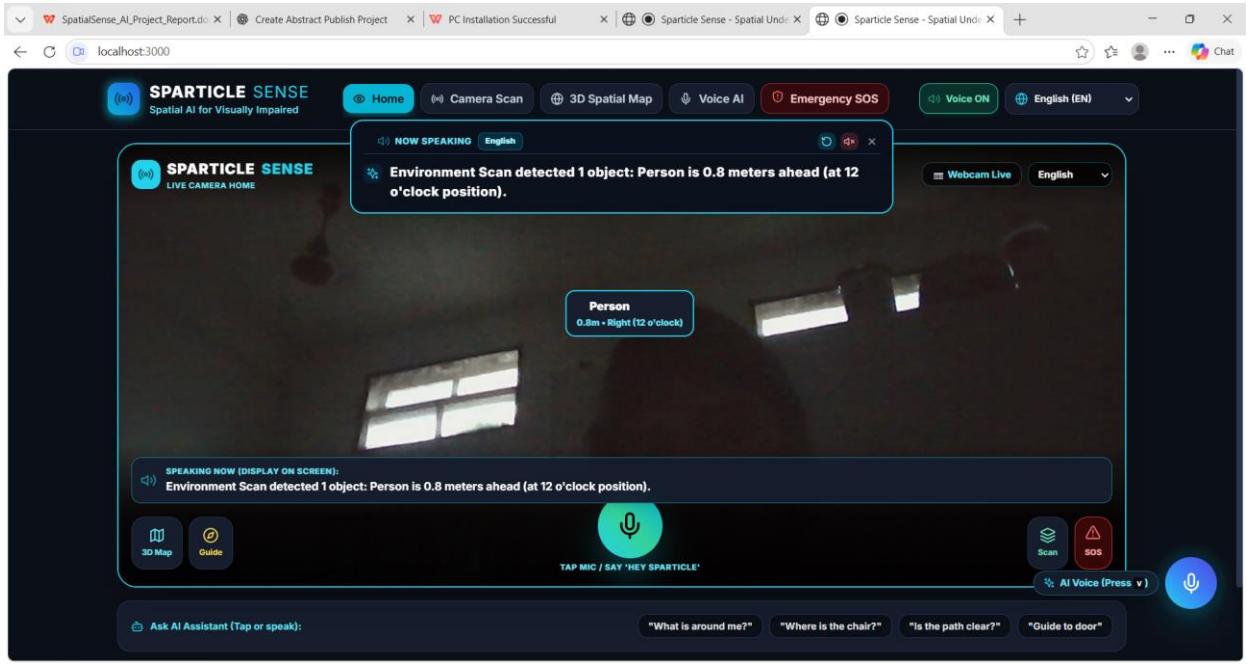


Fig. 2: Application home screen

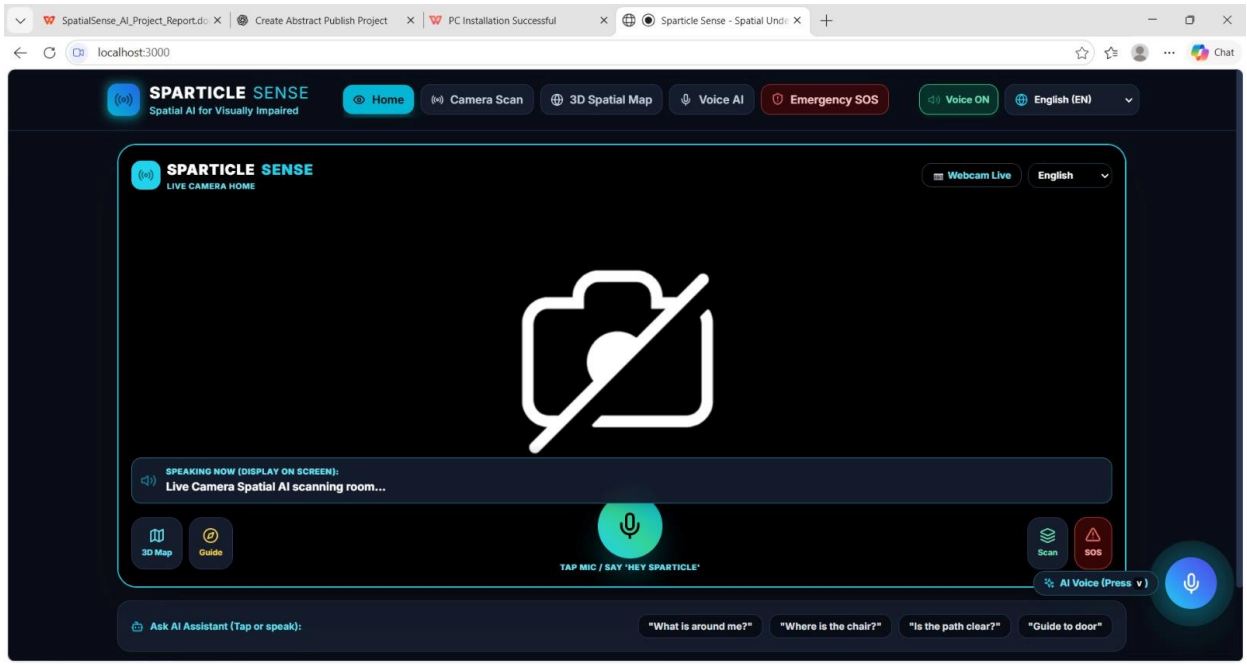


Fig. 3: Live camera detection with direction and distance

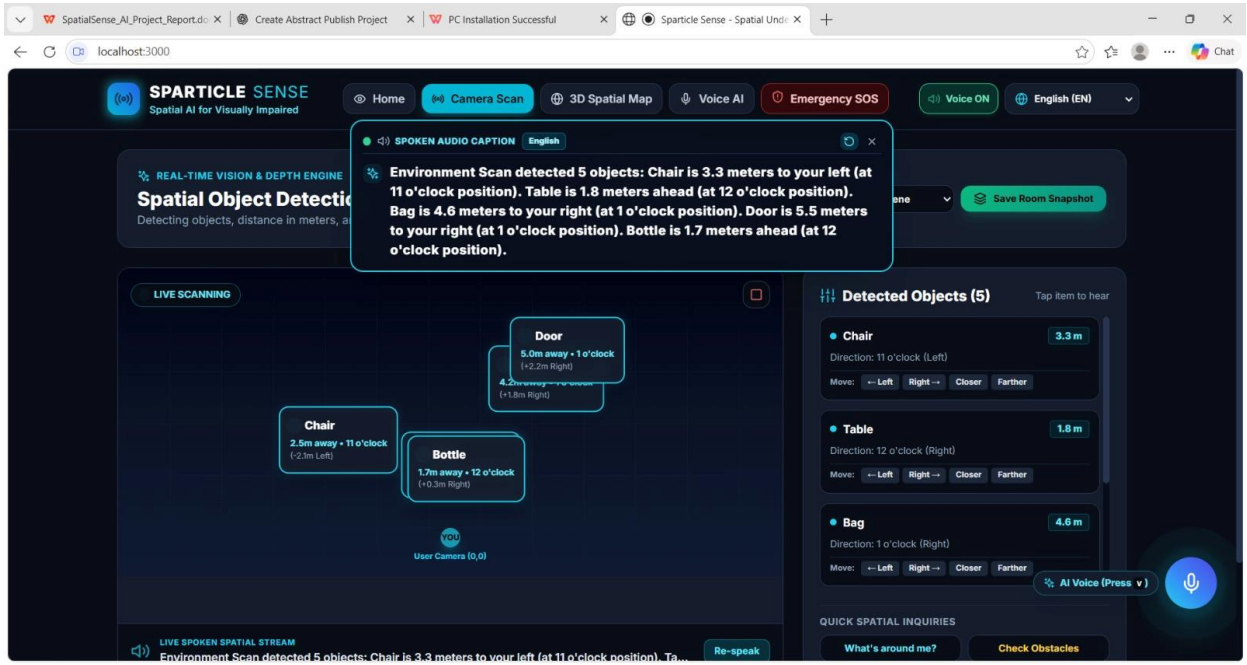


Fig. 4: Tracked objects with persistent IDs

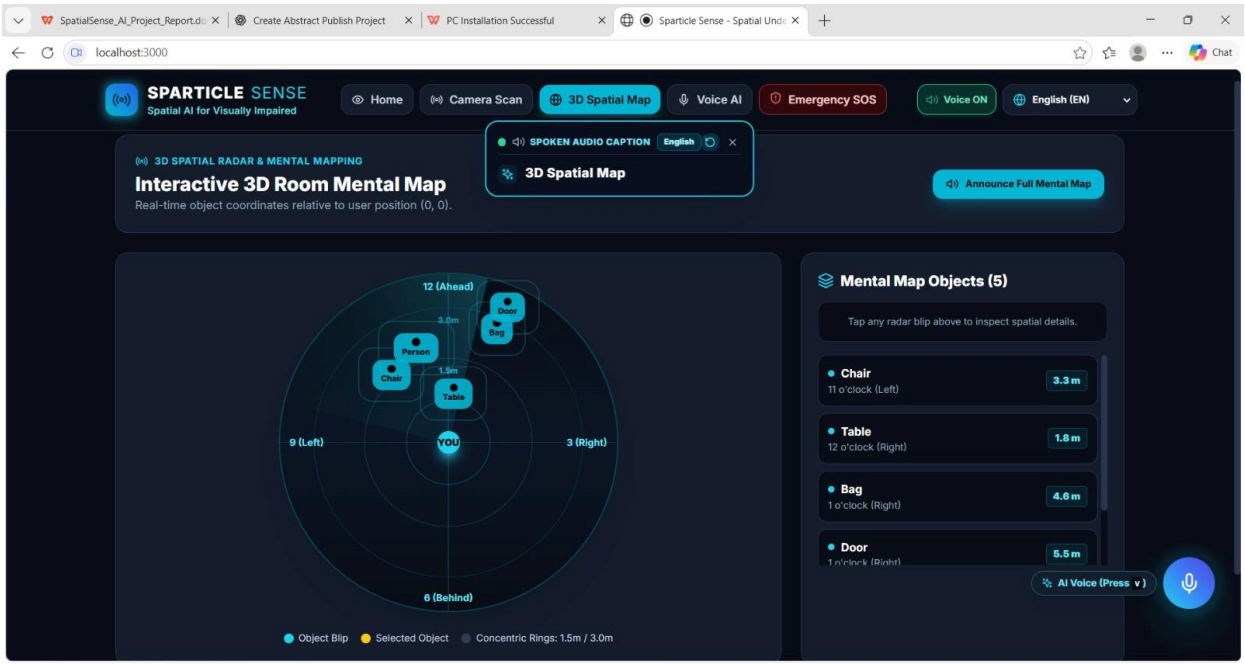


Fig. 5: Scene memory and change report (new / moved / missing)

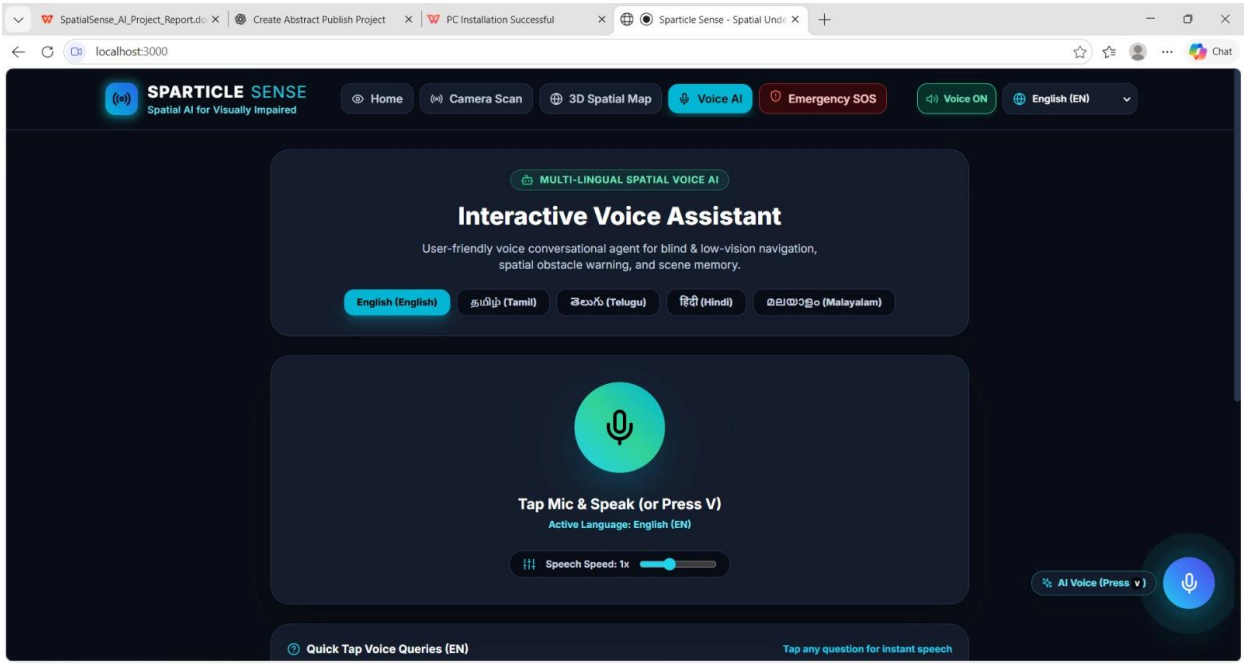


Fig. 6: Voice assistant question and spoken answer

LIMITATIONS AND SAFETY CONSIDERATIONS:

- Distance from a single camera is approximate and reported only as near, medium or far.
- Detection accuracy drops in low light, motion blur and clutter.
- Objects outside the camera view cannot be detected, so “missing” is never a guarantee of absence.
- The language model may phrase answers imperfectly, so responses are short and grounded in detections.
- The system is an assistive aid and must not be the only means of avoiding hazards such as stairs, traffic or drops.
- Validation with visually impaired users is required before any deployment claim.

CONCLUSION:

SpatialSense AI (EchoMap) shows that combining detection, tracking, spatial mapping and persistent memory can give visually impaired users a richer understanding of their surroundings than single-frame object detection. By announcing what changed, saying where things are, and stating clearly what is live and what is remembered, the system aims to be a practical and trustworthy assistive tool.

Future work includes calibrated depth estimation using depth models or sensors, full 3D room reconstruction, warnings for stairs and drops, GPS-based outdoor navigation, on-device inference to reduce latency and internet dependence, more Indian languages, and structured usability trials with visually impaired users and orientation-and-mobility experts.

REFERENCES:

[1] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[2] Jocher, G., Chaurasia, A., & Qiu, J. Ultralytics YOLOv8. Software, 2023. https://github.com/ultralytics/ultralytics

[3] Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. Simple Online and Realtime Tracking. Proc. IEEE International Conference on Image Processing (ICIP), 2016.

[4] Wojke, N., Bewley, A., & Paulus, D. Simple Online and Realtime Tracking with a Deep Association Metric. Proc. IEEE ICIP, 2017.

[5] Zhang, Y., Sun, P., Jiang, Y., et al. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. Proc. European Conference on Computer Vision (ECCV), 2022.

[6] Mur-Artal, R., Montiel, J. M. M., & Tardós, J. D. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. IEEE Transactions on Robotics, 31(5), 2015.

[7] Cadena, C., Carlone, L., Carrillo, H., et al. Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age. IEEE Transactions on Robotics, 32(6), 2016.

[8] Kuriakose, B., Shrestha, R., & Sandnes, F. E. Tools and Technologies for Blind and Visually Impaired Navigation Support: A Review. IETE Technical Review, 39(1), 2022.

[9] Messaoudi, M. D., Menelas, B.-A. J., & Mcheick, H. Review of Navigation Assistive Tools and Technologies for the Visually Impaired. Sensors, 22(20), 2022.