Anusandhanvallari

Vol 2025, No.1

December 2025

ISSN 2229-3388

AIVA – Artificial Intelligence Vision Assistant for Visually Impaired using YOLO and MiDaS Depth Estimation

Dr. Usha G R [1], Keerthi S [2], Prarthana G Kerudi [2], Pruthvi Raj B Y [2], Srujan M Amate [2]

ushagr85@gmail.com [1] keerthishivanna04@gmail.com [2] prarthangkerudi4239@gmail.com [2]

pruthvirajj084@gmail.com [2] srujanamate2004@gmail.com [2]

Associate Professor, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davanagere, Karnataka, India [1]

U.G Student, Department of Computer Science and Engineering,

Bapuji Institute of Engineering and Technology, Davanagere, Karnataka, India [2]

ABSTRACT

Individuals with visual impairments often face challenges from limited awareness of their surroundings. This affects their independence and safety in daily life. Our study presents a mobile based Artificial Intelligence Vision Assistant for people with visual impairments. It gives real time scene understanding through audio feedback. The system uses a smartphone camera to grab live video feeds. It relies on lightweight deep learning models like YOLO for quick object detection. EffiecientDet-Lite2 handles feature extraction well. TensorFlow Lite runs inference right on the device. The setup spots objects and picks out obstacles. It estimates distances too. It recognizes known people with a built in face recognition part. Visual details turn into natural speech via Android Text to Speech. This lets users get steady guidance without needing internet. The approach offers low delay and strong privacy. It delivers reliable offline work. All this boosts navigation and awareness for visually impaired folks. It helps their autonomy as well.

KEYWORDS

Artificial Intelligence Vision Assistant, Object Detection, EffiecientDet-Lite2, TensorFlow Lite, Face Recognition, Text-to-Speech, Assistive Technology.

1. Introduction

People with visual impairments face severe difficulties in terms of sensing their surroundings. Many activities that are part of everyday functioning become complex for them. Even locomotion without bumping into something assumes added anxiety. There are tools to read on a screen and apps to guide people. Most of these are not fully adapted to the present moment. They omit details of the surroundings. The user cannot consider everything that is around them with clarity. Such omissions seriously limit freedom of mobility that people with blindness deserve. [1]

Recent development in artificial intelligence and multi-modal systems has started to change that situation. Tools like Be My AI take advantage of advanced vision and language capabilities, delivering more precise descriptions of unfolding scenes to help with the understanding of locations and interactions with others present. The aids also have some problems: Their responses sometimes stray from the specific user request. They are overly dependent on a steady internet connection. [3]

That means there should be an evident need for better solutions. People would like to have assistants that can respond instantly, capturing the entire environment without blind spots. Such assistants should be really efficient on mobile phones and without bugs or lags. In conclusion, AIVA's objective is assistance that is truly accessible and corresponds to real user needs.

2. Problem Statement

People with visual impairments often struggle to understand their surroundings on their own. They put in a lot of mental work just to detect objects close by. Assistive technology has improved quite a bit over the last few years. Even so, the gap in cost between high-end devices and simpler ones stays pretty wide. Many of these tools end up being too bulky or heavy to carry around daily. Such designs limit easy movement in routine tasks. They also slow down the immediate help that users depend on. People who are blind obviously require a low-cost, lightweight gadget that pairs with smartphones. It needs to identify and describe visual environments right away. The device ought to give reliable details on how far away nearby objects are. Options like that would improve safety for people dealing with vision challenges. They might gain more confidence in navigating various places on their own.

3. Scope and Methodology

3.1 Scope

The project presented here is an integrated Artificial Intelligence Vision Assistant, by the name AIVA. It is a mobile system that offers real-time audio guidance for people dealing with visual impairments. According to available studies, individuals with visual challenges benefit most from tools that combine object detection, text reading, face recognition and navigation help within one package. Current solutions have a tendency to separate these features. They also frequently lack real contextual understanding. [6]

AIVA unifies major AI functions that are shown to be beneficial in existing tools. These include real- time object detection, obstacle detection, face identification and text recognition via OCR. Related work also identifies the importance of mixing computer vision with simple audio or voice-based interactions to improve accessibility, ease manipulation and encourage independence for people with low vision. [3]

The project adds in voice commands along with organized audio responses. This addresses the move for user-friendly designs seen in lately developed navigation aids and IoT supports aimed at the visually impaired. AIVA handles failed detections, commands not registering and faces that it cannot place. Such problems are very common in most existing tools. Finally, the system runs and is tested on Android phones. This aligns well with findings favoring mobile setups offering quick responses and real-world practicality in boosting safety and self-reliance.

In a nutshell, the initiative aims to increase independence, awareness of the environment and safety in general. This it does by converting challenging visual inputs into quick, helpful audio prompts for the visually impaired.

3.2 Methodology

The methodology describes the process of capturing visual and voice input through AIVA, processes it through multiple AI components and converts the result into clear, real-time audio guidance. It outlines the flow of data, the interaction between modules and the steps taken to ensure that the support is accurate, fast and user-friendly for Visually impaired people.

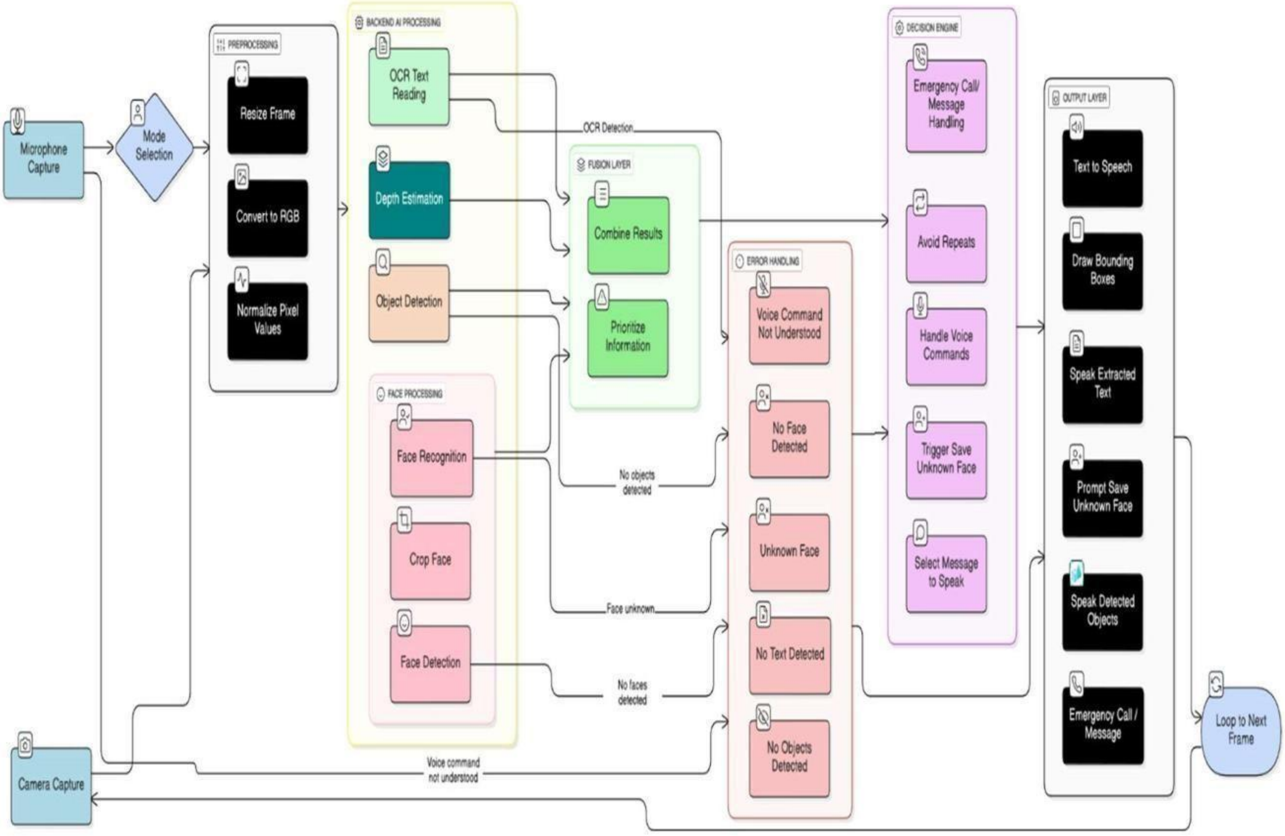


Figure 3.2.1 : System Workflow of the Proposed AIVA Model

1. **Input Acquisition**

   AIVA continuously captures video from a smartphone camera while listening for voice commands. Hands-free, speech-based interaction is widely regarded as an attribute of accessibility and ease of use for assistive AI systems.

2. **Mode Selection**

   Depending on user voice input, the system triggers the appropriate mode: object detection, face recognition, OCR or scene description. This modular task switching mirrors the approaches used in tools like Seeing AI, which employ multiple "channels" for different tasks.

3. **Mode Selection**

   Input frames are resized, normalized and converted to RGB before AI inference. Preprocessing enhances the model accuracy, similar to approaches currently used in state-of-the-art OCR and MathML conversion pipelines relying on structured images preparation. The detected faces are cropped in order to improve recognition reliability.

4. **Core AI Processing**

Face Detection & Recognition

By identifying faces and comparing these with stored profiles, it supports social interaction- an ability considered valuable in modern assistive devices such as OrCam MyEye.

Text Recognition (OCR)

OCR locates text and reads it out loud, consistent with the principles of text accessibility used in Seeing AI and Benetech's accessible document processes

Object Detection

Lightweight models, like YOLO or EfficientDet-Lite2, perform real-time object detection on mobile devices. This is in line with the growing trends of mobile assistive AI systems.

Depth Estimation

Distance measurement has helped users understand their surroundings, thus enabling safer mobility a key element. Highlighted in assistive navigation research.

5. Integration and Fusion

The outputs from all modules are combined, giving precedence to nearby obstacles or significant objects. Research Emphasizes the importance of filtering and organizing information to avoid overwhelming visually impaired users.

6. Error Handling

The system manages unclear commands, missing detections and unknown faces while providing corrective audio prompts. According to studies, feedback like this increases usability in assistive technologies greatly.

7. Decision Logic & Output

Decision modules select the most salient information, avoid repetition, manage tasks that involve saving new faces. Finally, the results are delivered through text-to-speech, similar to how operational methods work in widely used tools like OrCam, Seeing AI.

4. Literature Review

Assistive technologies such as vision, language, and wearability solutions for visually impaired individuals have been improved by AI, but issues such as cost, privacy, and lack of adaptation pose challenges, thereby indicating a need for more reliable solutions. [1]

AI-powered assistive technologies facilitate navigation and object recognition in visually impaired persons, although most of the systems still face challenges related to issues of lighting, limited datasets, and multilingual aspects. Ongoing enhancements are made to achieve better real-time perception and independence. [2]

Large multimodal models, such as GPT-4, improve the interpretation of scenes and interaction with users by providing richer support for visually impaired users. However, problems like hallucinations and limited goal-directed responses show that more reliable systems must be developed. [3] SLAM- based IoT systems like Blindaid help locate objects and support indoor navigation using real- time mapping and audio feedback. Their low-cost design enhances accessibility and practical usability for visually impaired users. [4]

AI tools are revolutionizing accessibility in healthcare, education, and day-to-day living with personalized support. However, privacy, algorithmic bias, and costs remain major concerns for broad adoption. [5] Existing image-recognition tools lack sufficient accuracy, contextual clarity, and usability for blind users. The study emphasizes inclusive design guided by ISO and multimodal interaction for better support of diverse user needs. [6] Large multimodal assistants, like Be My AI, offer richer scene descriptions and support daily tasks for visually impaired users. However, challenges such as hallucinations and limited goal-oriented responses show the need for more reliable and context-aware assistive AI systems. [7] YOLO-based object detection combined with monocular depth estimation and spatial audio improves indoor awareness for visually impaired users, but high computational requirements and hardware dependency limit real-time portability. [8]

The study accentuates the role of AI-based vision techniques in providing supporting navigation and object awareness for visually impaired users; however, it remains limited to a mere promise in practical deployment due to accuracy issues and variability in the real world. [9] The object detection system with the text-to-speech alert mechanism based on YOLO is a real-time improvement in object awareness for the visually impaired, but it is a function of the available dataset size and computational power. [10] AI-based vision and voice assistance enhances object awareness and independence among users with poor vision, although real-time accuracy and hardware limitations are still key challenges to be met in practical applications. [11]

The AI-driven smart navigation with real-time object detection and audio feedback empowers VI users with better environmental awareness and independent mobility, although further studies indicate that performance may be subject to environmental complexity, device limitations, and reliance on mobile hardware. [12] AI-powered mobile applications using YOLO and audio outputing capability enhance object detection and navigation support for the visually impaired, but the challenges of real-time recall capability and resource utilization of smartphone power remain major drawbacks. [13] A mobile edge AI solution for visual assistance using stereo vision, optimized deep learning models, and low-power edge hardware for real-time obstacle detection for visually impaired individuals. It highlights enhanced mobility and feasibility for edge deployment of visually aided navigation systems.

[14] A YOLO-based indoor navigation system uses monocular depth estimation and binaural spatial audio to convey obstacle location and distance for visually impaired users, demonstrating effective real-time vision-driven audio assistance. [15]

5. Result and Discussion

Our AIVA system pulls together a few key computer vision tools into a single Android app. We designed it just for people who are visually impaired. We put in YOLOv8-Nano Float16 to handle object detection. It runs pretty smoothly on phone hardware. Inference times sit between 18 and 30 milliseconds for each frame. That kind of speed lets the system give real-time audio cues about objects it spots. It also tells users where those objects sit, like left, center or right from their view.

We went with ML Kit's OCR setup for recognizing text. It does a solid job on printed stuff. Handwritten text still trips it up sometimes. We added a voice prompt right before the app reads any text out loud. This setup stops those pesky interruptions. It puts control in the users' hands when they want to hear the text.

The face recognition part holds up well under normal lighting. Users save faces with a simple voice okay. The system relies on an embedding method to recall people over multiple uses. We brought in Depth Anything V2 Large to estimate depths.This helps users get a rough idea of distances to things when they request it. We created a separate thread for inference. We also added caching to keep the depth model from slowing everything down. Testing in the field proved AIVA acts as a useful helper tool. The wake-word spotting, speech input and voice output build an easy interface.

The following snapshots present the system's real-world outputs and interface interactions, illustrating how AIVA performs during practical usage.

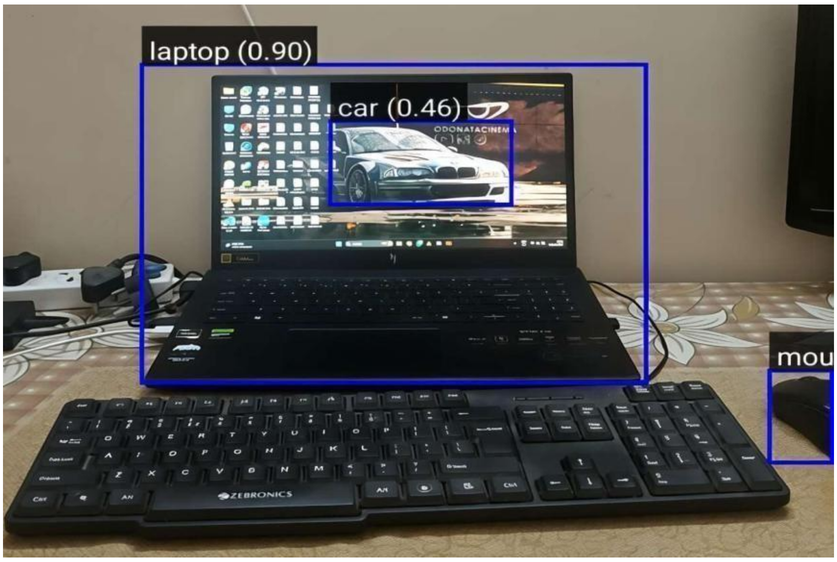


Figure 5.1: Object Detection

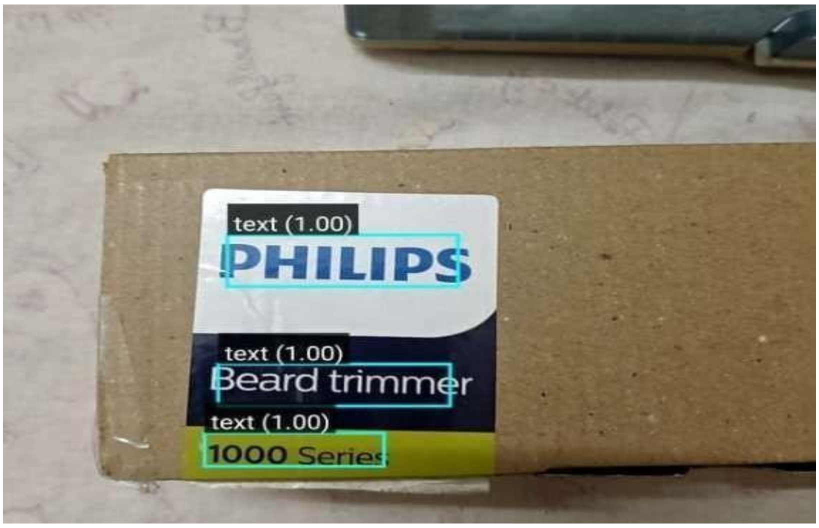


Figure 5.2: Text Recognition

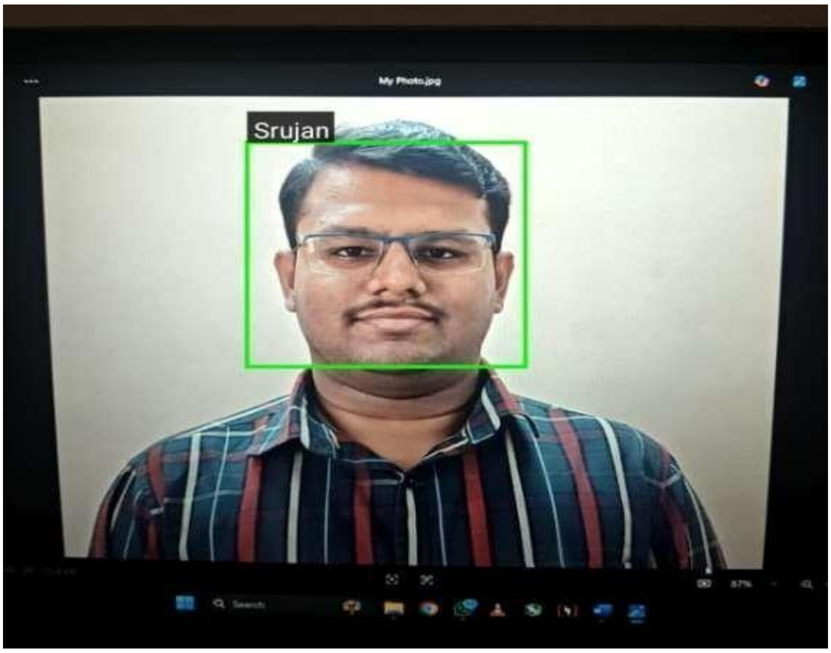


Figure 5.3: Face Recognition

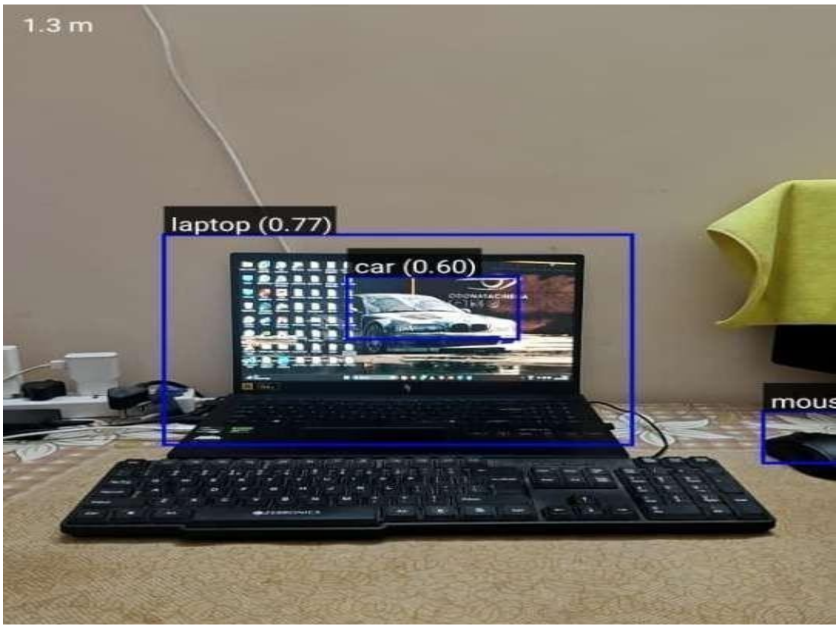
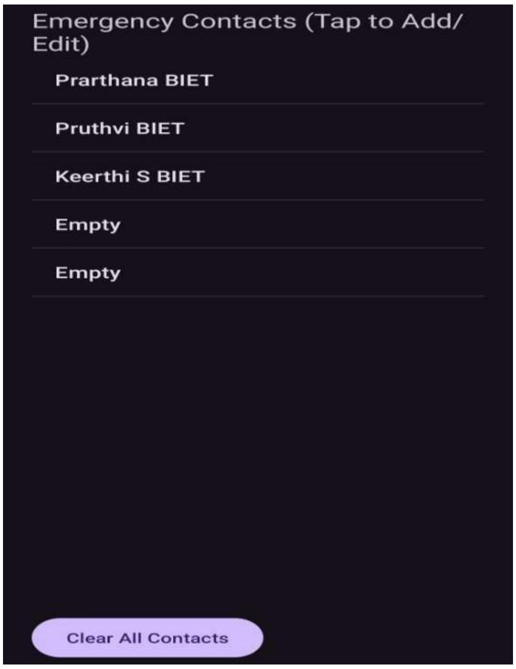


Figure 5.4: Depth Estimation



Emergency Contacts (Tap to Add/ Edit)

Prarthana BIET

Pruthvi BIET

Keerthi S BIET

Empty

Empty

Figure 5.5: Emergency Contacts

6. Findings

We picked up some solid lessons during the build and tests.

- Lightweight models really count when deploying on phones. YOLOv8-Nano and EfficientDet- Lite2 both did fine. EfficentDet-Lite-2 struck the best balance of accuracy against speed.
- Putting object detection, text reading, and face recognition all in one app turned out more helpful than scattered tools. Users liked skipping the hassle of app switching.
- Hands-free voice control matters a lot for these users. We mixed Porcupine for wake words, Android's and text-to-speech. That combo shaped a strong way to interact.
- Depth estimation brought real safety value. It lets users judge distances to close obstacles and items.
- We set up scanning that picks frames based on context. It sends them only to the right models. This cut down on processing demands. It kept things from lagging in real time.

7. Limitations and Research

AIVA does well in most ways. Still, a few weak spots need work.

- Detection accuracy is severely compromised under low or uneven light conditions, making it quite challenging for the system to correctly identify objects or text. This limits performance in night-time or low-light indoor environments.
- Background noise can interfere with wake-word detection and disrupt speech commands, reducing the system’s

responsiveness. This makes voice-based interaction difficult in crowded or noisy environment.

- Continuous shooting with cameras and real-time AI processing rapidly drain the battery, reducing the operating time of the system. This will limit long-duration use and require more frequent charging.

8. Conclusion

AIVA proves we can blend AI vision pieces like object detection, text recognition, face recognition, and depth estimation into one helper for visually impaired users. The app delivers instant feedback. It runs hands-free via voice. It aids in grasping surroundings. Tests show it moves quick and true on everyday Android phones. This version hits our core aims. Room exists for tweaks, though. Sharper depth reads, stronger low-light handling, power savings and extra toughness would lift AIVA higher. The work marks real steps in AI tools that aid daily life. They boost freedom and caution for those without sight.

References

[1] P. Naayini, "AI-Powered Assistive Technologies for Visual Impairment," *arXiv,* p. 12, 2025.

[2] M. Kumar, "AI ASSISTANT FOR VISUALLY IMPAIRED," *International Journal of Trendy Research in Engineering and Technology,* vol. 9, no. 4, p. 12, 2025.

[3] J. Xie, "Emerging Practices for large multimodal model (LMM) Assistance for people with visual impairments: Implications for Design," *Proceedings of the ACM Conference on Human Factors in Computing Systems,* p. 22, 2015.

[4] D. S. G, "Blindaid: Assisting The Visually Impaired In Object Detection and Tracking Using Slam," *International Journal of Environmental Sciences,* vol. 11, p. 12, 2025.

[5] S. Ferebee, "AI and Accessibility: Breaking Barriers for People with Disabilities," *PREMIER JOURNAL OF ARTIFICIAL INTELLIGENCE,* p. 10, 2025.

[6] V. S. FERNANDO, "Image Recognition Tools for Blind and Visually Impaired Users: An Emphasis on the Design Considerations," *ACM Transactions on Accessible Computing,* vol. 18, no. 1, p. 21, 2025.

[7] J. Xie, "Beyond Visual Perception: Insights from Smartphone Interaction of Visually Impaired Users with Large Multimodal Models," *CHI, Conference on Human Factors in Computing Systems,* p. 17, 2025.

[8] A. M. George, "YOLO-Based Object Recognition System for Visually Impaired," *International Journal of Science and Engineering Applications (IJSEA),* vol. 14, no. 1, p. 9, 2025.

[9] R. Gonzalez, "Investigating Use Cases of AI-Powered Scene Description Applications for Blind and Low Vision People," *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24),* p. 21, 2024.

[10] D. M. Y. Babu, "Object Detection System with Voice Alert for Blind," *International Journal for Research in Applied Science & Engineering Technology (IJRASET),* vol. 11, no. 2, p. 5, 2023.

[11] K. S. D, "Deep Learning Virtual Assistant for Visually Impaired with Object Detection and Distance Estimation," *Grenze International Journal of Engineering and Technology,* vol. 9, no. 2, p. 8, 2023.

[12] A. P. Rajvardhan Shendge, "Smart Navigation for Visually Impaired People Using Artificial Intelligence," *ITM Web of Conferences,* vol. 44, p. 7, 2022.

[13] T. A. Qureshi, "AI Based App for Blind People," *International Research Journal of Engineering and Technology,* vol. 8, no. 3, p. 5, 2021.

[14] J. K. Mahendran, "Computer Vision-Based Assistance System for the Visually Impaired Using Mobile Edge Artificial Intelligence," *IEEE Access ,* vol. 9, p. 16, 2021.

[15] X. Y. Sukesh Davanthapuram, "Visually Impaired Indoor Navigation using YOLO-Based Object Recognition, Monocular Depth Estimation and Binaural Sounds," *Proceedings of the IEEE International Conference on Electro/Information Technology,* p. 10, 2021.