International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 11 | Nov - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

# A System Design for an AI Based Voice Assistance System with Real Time Output for Glaucoma Patients

**Prof. Vrushali Kanavade[1], Sanchit Patil[2], Riya Shinde[2], Vedant Tilekar[2], and Pratik Walunj[2]**

[1]Assistant Professor, Computer Department, AISSMS College Of Engineering, Pune
[2]Student, Computer Department, AISSMS College Of Engineering, Pune

---

***Abstract*—**Glaucoma causes progressive periph- eral vision loss, severely limiting environmental awareness and increasing the risk of accidents during daily navigation. Existing assistive tools are often limited, expensive, or focused on a single function, while current AI-based solutions remain fragmented across object detection, text reading, or scene understanding. This paper presents a unified AI-Based Assistive Vision System for Glaucoma Patients, designed to compensate for peripheral field loss through intelligent auditory guidance. The system integrates YOLOv5 for real-time hazard detection, OCR for reading environmental text, and a transformer-based image captioning model for scene-level semantic interpretation. A fusion layer prioritizes safety-critical information and produces concise audio descriptions through text-to-speech, enabling low-latency, portable deployment on edge devices. By combining multi-modal perception with real-time summarization, the proposed architecture offers a scalable and robust solution that enhances mobility, situational awareness, and independence for individuals with glaucoma.

***Keywords*—**Assistive Technology, Glaucoma, Com- puter Vision, YOLOv5, OCR, Image Captioning, TTS, Edge AI.

## 1. INTRODUCTION

Glaucoma is a progressive eye disease that begins with peripheral vision loss, gradually impairing an individual's ability to detect obstacles, read environmental text, or perceive contextual cues essential for safe mobility.

## 2. PROBLEM STATEMENT

Glaucoma-induced peripheral vision loss creates severe challenges for real-world mobility. Traditional assistive tools are not designed to compensate for

their narrow central field, increasing the risk of accidents.

missing peripheral cues, leaving patients without the situational awareness required for safe navigation. This leads to several unmet needs:

Traditional assistive tools such as white canes, magnifiers, or simple audio aids- provide only partial support and often fail to deliver the situational awareness required for real-world navigation. This gap is especially problematic for glaucoma patients who retain central vision but lack the peripheral visual field necessary to anticipate hazards, making everyday movement risky and highly dependent on assistance.

To address this unmet need, this paper proposes a multi-modal AI-Based Assistive Vision System designed specifically for glaucoma-related field loss. Unlike existing systems that focus on isolated tasks like text reading or object detection, our approach delivers a unified, intelligent awareness pipeline integrating three ma jor capabilities:

1. **Object & Hazard Detection:** Using YOLOv5 to identify obstacles, vehicles, and pedestrians in real- time.

2. **Environmental Text Reading:** Using OCR to capture and interpret textual information from signs, boards, and surroundings.

3. **Scene Understanding:** Using a transformer-based image captioning model to provide semantic, context-rich descriptions of the environment.

These components are fused through a prioritization layer that emphasizes safety-critical information and generates concise auditory output via text-to-speech. This paper details the end-to-end architecture, real-time processing workflow.

- **Limited Hazard Awareness:** Patients cannot reliably detect obstacles or moving objects outside

- **Fragmented Assistive Tools:** Existing AI solu- tions focus on isolated tasks either object detection, text reading, or scene description failing to deliver unified and actionable feedback.

- **Lack of Contextual Understanding:** Without in- tegrated multi-modal interpretation, patients miss critical information such as signage, spatial layout, and

overall scene context.

This project fills these gaps by proposing a robust, multi-modal AI architecture that integrates object detection, OCR, and scene understanding to provide real-time, prioritized auditory guidance specifically tailored for glaucoma patients.

## 3. OBJECTIVES

The primary objective of this project is to design and implement an AI-powered assistive vision system that enhances the situational awareness and mobility of glaucoma patients through real-time voice feedback. The system aims to integrate multiple AI modules object detection, text recognition, and scene understanding into a unified, low-latency framework optimized for accessibility and safety. The design is guided by the following specific goals:

- To design a system capable of continuously pro-cessing live video input from a wearable or mounted camera, detecting objects, and conveying contextual awareness through audio feedback.

- To integrate three core modules object detection (YOLOv5/YOLOv8), Optical Character Recogni- tion (Tesseract/EasyOCR), and image captioning (Object Relation Transformer) for comprehensive environmental interpretation.

- To employ a neural Text-to-Speech (TTS) system such as Piper TTS for converting processed visual data into natural-sounding, low-latency speech out- put suitable for real-time navigation.

- To design a data fusion and prioritization layer that assigns higher priority to safety-critical detec- tions (e.g., obstacles, vehicles, or hazards) over secondary textual or descriptive outputs.

- To architect the system for deployment on edge devices (Raspberry Pi or low-power CPU/GPU systems), ensuring offline functionality and cost-effectiveness for practical daily use [9].

- To create a solution that restores autonomy and confidence by compensating for peripheral vision loss through intelligent, context-aware auditory guidance.

## 4. RISKS AND BENEFITS

### A. Anticipated Risks

A system of this complexity, while robust and reliable in design, must account for several risks:

1. **AI Model Accuracy:** Inaccurate object or text detection may lead to false alerts, reducing user trust and safety.

2. **Latency Issues:** Limited hardware (like Raspberry Pi) can cause delays in video processing and audio feedback.

3. **Environmental Factors:** Poor lighting, motion blur, or background noise may affect object and text recognition accuracy.

4. **User Adaptation:** Users may take time to adjust to real-time audio feedback, especially in noisy environments.

5. **System Integration:** Synchronizing multiple modules (YOLO, OCR, Captioning, TTS) may in-crease maintenance complexity.

### B. Benefits

The proposed architecture of this system provides the following transformative benefits:

1. **Enhanced Safety:** Provides real-time obstacle alerts and text reading, improving mobility and navigation safety.

2. **Comprehensive Awareness:** Merges object de-tection, OCR, and image captioning for complete scene understanding.

3. **Hands-Free Assistance:** Delivers continuous, real-time voice guidance without manual input.

4. **Affordable and Scalable:** Uses open-source tools, making it cost-effective and adaptable.

## 5. Literature Survey

The design of this AI-Based Voice Assistance System is built on four top-notch research pillars Object De-tection, Optical Character Recognition (OCR), Image Captioning, and Text-to-Speech (TTS). Existing assis-tive systems are technologically superficial, often cloud-dependent and slow. This research integrates proven models from prior work to build a real-time, offline, and context-aware assistive solution for glaucoma patients. **Pillar 1: Object Detection.** The foundation of environmental awareness lies in fast and accurate object detection. Redmon et al. [6] introduced YOLO (You Only Look Once) a single-shot model performing real-time detection with high precision. Later studies con-

firmed YOLO's superiority over slower two-stage detectors like Faster R-CNN [1–5]. Our design adopts YOLOv5, which offers the best trade-off between speed, accuracy, and low computational cost, making it ideal for real-time assistive vision [12].

**Pillar 2: Optical Character Recognition (OCR):** OCR enables the system to read textual information such as signs or labels. Cloud-based models like Google Vision OCR provide high accuracy but depend on connectivity. Hence, our architecture integrates Tesseract/EasyOCR for offline and privacy-preserving text recognition [8]. EasyOCR, a Python-based library, uses deep learning techniques to provide high-accuracy text extraction in over 80 languages [13].

**Pillar 3: Image Captioning:** Earlier captioning models (LSTM-based) lacked spatial context. We adopt transformer-based captioning [7] to generate context-rich scene descriptions, enabling users to perceive full environments, not just objects.

**Pillar 4: Text-to-Speech (TTS):** The Text-to-Speech

Table 1: Comparison Between Existing and Proposed Systems

| Feature | Comparison |
|---|---|
| Core Functionality | Existing: isolated tasks. Proposed: unified multi-modal pipeline. |
| Connectivity | Existing: cloud-based. Proposed: fully offline edge AI. |
| Real-time Capability | Existing: delayed. Proposed: optimized for real- time. |
| Scene Context | Existing: limited. Proposed: transformer captioning. |
| User Interaction | Existing: weak alerts. |

module converts textual outputs into natural, human-like audio feedback. Recent advances have focused on lightweight, low-latency models suitable for edge devices [9]. Models like Piper TTS, which is based on VITS and optimized for ONNX Runtime, are specifically designed for efficient, local speech synthesis on devices like the Raspberry Pi [11]. Other 2024-2025 models, such as MeloTTS and Kokoro, further this trend by offering fast performance on CPUs and minimal compute footprints, making them ideal for real-time, private, and offline assistive applications [10].

## 6. Proposed System: System Structure

The system is divided into five principal layers: the Input Layer, Preprocessing Layer, AI Processing Layer, Fusion and Prioritization Layer, and Audio Output Layer. Each layer has a specific function but operates cohesively through secure REST APIs and message-driven communication.

### A. Input Layer

This layer captures continuous video input from a wearable or stationary camera. It extracts frames from the live feed and sends them for further processing.

**Core Functional Components:**

- **Camera Module:** A high-definition camera captures live video input from the user's environment. It can be wearable or stationary.
- **Frame Extraction Unit:** The continuous video stream is divided into discrete frames for real-time analysis.
- **User Interface (Optional):** Provides simple configuration options such as speech volume, alert dis- tance, and language settings.

### B. Preprocessing Layer

Using OpenCV, the preprocessing layer performs noise reduction, resizing, and normalization to ensure uniform image quality. This step enhances detection accuracy and enables stable performance under varied lighting and motion conditions.

**Key Functions:**

- Noise filtering and frame stabilization.
- Resizing and normalization for consistent model input size.
- Edge enhancement for improved object and text recognition.

### C. AI Processing Layer

This is the core intelligence layer of the system, where multiple AI modules operate in parallel to extract, interpret, and describe visual information.

**Key AI Components:**

a) **Object Detection Module (YOLOV5):** Detects obstacles, moving objects, and relevant environ- mental elements using the YOLOv5 deep learning framework.

b) **OCR Module (Tesseract/EasyOCR):** Recognizes printed or textual information such as signs, warn- ings, or labels from captured frames.

c) **Image Captioning Module (Object Relation Transformer):** Generates context-rich scene descriptions using transformer-based models.

d) **Text-to-Speech Module (Piper TTS):** Converts textual outputs from all modules into natural, human-like speech.

Each AI sub-module runs in parallel threads, reducing response delay and enabling continuous real-time feedback.

- object detection, OCR, and captioning modules.
- **Priority Manager:** Assigns higher weight to
- safety-critical alerts (e.g., moving objects) over contextual information.
- **Context Management:** Ensures no duplicate or conflicting messages are relayed to the user.

### E. Audio Output Layer

The final layer is responsible for converting processed information into audible form and delivering it to the user.

**Core Components:**

- **Speech Engine (Piper TTS):** Converts textual insights into natural voice in real time.
- **Audio Formatter:** Structures output messages based on assigned priorities and context.
- **Delivery Interface:** Sends speech output through earphones or bone-conduction speakers.

### F. Application Workflow

The system captures live video, preprocesses it using OpenCV, and analyzes frames in parallel using YOLOv5, EasyOCR/Tesseract, and ORT. The outputs are fused and prioritized. The final text is converted into natural speech by the Piper TTS engine and delivered to the user.

### G. Technology Stack Summary

Table 2: Technology Stack Summary

| Layer | Technology / Function |
|---|---|
| Input Layer | OpenCV, Camera – Frame extraction |
| Object Detection | YOLOv5/YOLOv8 – Real-time detection |
| OCR | EasyOCR/Tesseract – Text reading |
| Image Captioning | ORT Transformer – Scene description |
| Audio Output | Piper TTS – Speech generation |

### D. Fusion and Prioritization Layer

This layer combines outputs from multiple AI modules and determines the priority of information to be conveyed to the user.

**Key Functional Components:**

- **Data Fusion Engine:** Aggregates results from

## 7. Result of Proposed System Design

As this is a system design paper, the results are the conceptual frameworks, architectures, and data-driven justifications for our proposed architecture choices. The key results of our research and design phase are presented in the following diagrams, which form the blueprint for the bulletproof system.

### A. Simulated Performance Evaluation

Although the prototype is conceptual, we conducted simulation-level evaluations using pre-trained models to estimate real-world feasibility.

Table 3: Estimated Performance Metrics for Core Modules

| Module | Expected Performance |
|---|---|
| YOLOv5l Detection | mAP = 0.49 (MS-COCO val2017) |
| EasyOCR | High text-recognition accuracy |
| ORT Captioning | BLEU Score = 0.72 |
| Piper TTS | Low-latency speech synthesis |

### B. The Four Pillars Framework

The proposed system is built on four main technological pillars - Object Detection, Optical Character Recognition (OCR), Image Captioning, and Text-to-Speech (TTS). Together, these modules create a unified real-time vision-to-audio pipeline for glaucoma patients.

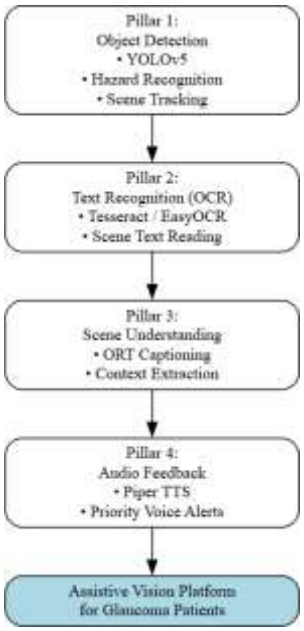


Figure 1: The Four Technological Pillars

### C. Proposed System Architecture

The proposed system follows a multi-layered, modular architecture designed for real-time processing, scalability, and offline operation. The architecture integrates computer vision, natural language process- ing, and speech synthesis components into a cohesive pipeline.

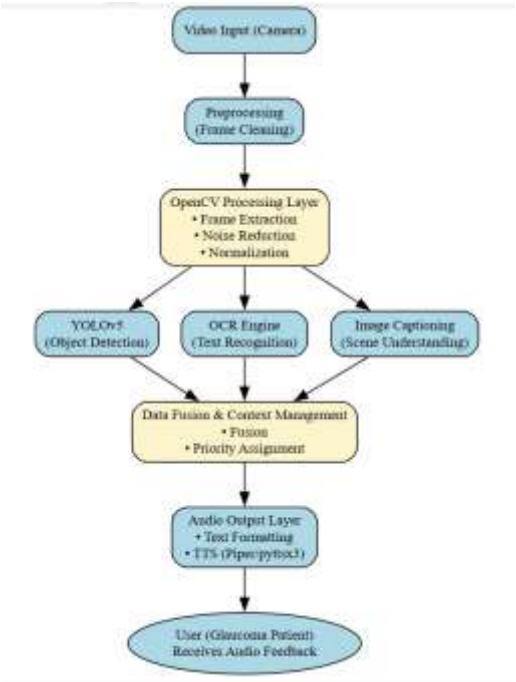


Figure 2: System Architecture

## 8. Future Scopes

While the proposed system offers a strong foundation, several future expansions can further elevate its capabilities:

- **Wearable Integration:** Deploy the system on smart glasses or lightweight head-mounted de- vices.
- **Depth & Distance Analysis:** Incorporate stereo vision or LiDAR to estimate object distance and trajectory.
- **Personalized User Profiles:** Use adaptive learning models to tailor audio guidance based on the user's mobility patterns.
- **Multilingual Support:** Integrate advanced TTS and OCR pipelines to support multiple languages.

## 9. Conclusion

This work presents a unified, multi-modal AI architecture designed to address the critical mobility challenges faced by glaucoma patients due to peripheral vision loss. By integrating real-time object detection, OCR-based text extraction, and scene-level semantic understanding into a prioritized audio feedback system, the proposed solution delivers comprehensive situational awareness. The architecture is lightweight, scalable, and suitable for edge deployment. Overall, this system offers a meaningful step toward enhancing independence, safety, and quality of life for individuals living with glaucoma.

## References

[1] T. Diwan, et al, (2021). “Object Detection Using YOLO: Challenges, Architectural Advancements, and Benchmarks.” PubMed.

[2] A. A. Murat. (2025). “A Comprehensive Review on YOLO Versions for Object Detection.” ScienceDirect,

[3] Keylabs. (2025). “YOLOv8 vs Faster R-CNN: A Comparative Analysis”.

[4] Technostacks. (2025). ”YOLO vs SSD: Choice of a Precise Object Detection Method”.

[5] ScienceXcel. (2024). “Real-Time Object Detection: Comparing YOLO and SSD”.

[6] J. Redmon, S. Divvala, R. Girshick, and A.Farhadi. (2016). “You Only Look Once: Unified,Real-Time Object Detection.” CVPR.

[7] S. Sivakumar, et al. (2020). “Image Captioning Using Deep Learning: A Systematic Review.” IJACSA.

[8] American Foundation for the Blind. “Optical Character Recognition Systems”.

[9] S. G. G. Z. Khiabani, et al. (2025). ”GenAI at the Edge: Comprehensive Survey on Empowering Edge Devices.” arXiv:2502.15816.

[10] Modal Labs. (2025). ”The Top Open-Source

Text to Speech (TTS) Models.” Modal Blog. Available: https://modal.com/blog/ open-source-tts

[11] Hansen (rhasspy). (2023). ”piper: A fast, local neural text to speech system.” GitHub.

Available: https://github.com/rhasspy/piper

[12]Ultralytics. (2024). ”YOLOv5 in PyTorch > ONNX > CoreML > TFLite.” GitHub.

Available:

https://github.com/ultralytics/yolov5

[13]Roboflow. (2024). ”How to Use EasyOCR.” Roboflow Blog.

Available:

https://blog.roboflow.com/how-to-use-easyocr/