iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82515

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

A Survey on Hybrid Approaches for Real-Time Visual Perception in Assistive Systems

Pranav Gupta[1], Bharati Kumari[2], Anjali Singh[3], Shylendra S[4], Dr. Kamalakshi Naganna[5]
[1, 2, 3, 5]*Department of Computer Science and Engineering Sapthagiri College of Engineering, Bengaluru, Karnataka, India*
[5]*Professor & Head of Department Department of Computer Science and Engineering, Sapthagiri College of Engineering Bengaluru, Karnataka, India*

Abstract: ***Visual impairment is a significant disability that restricts an individual's ability to perceive, navigate, and interact with the surrounding environment independently. The rapid advancement of deep learning and computer vision has led to a growing body of research on AI-assisted perception systems for visually impaired users. However, existing studies predominantly rely on single-model object detection pipelines, which exhibit an inherent trade-off between detection accuracy and inference speed. This survey reviews ten recent works in the domain of assistive visual perception systems (compared in Table I), with a focus on object detection architectures, multimodal integration strategies, and edge deployment approaches. The reviewed works encompass a variety of methods, including YOLO-based detectors (YOLOv5 through YOLOv11), transformer-based detectors such as RF-DETR Nano, Optical Character Recognition (OCR), Speech-to-Text (STT), and Text-to-Speech (TTS) modules, as well as dedicated edge AI accelerators such as the Google Coral Edge TPU and augmented reality platforms such as the Vuzix Blade 2. The survey examines the strengths and limitations of each approach and identifies a recurring research gap in the adaptive, context-aware selection of detection models based on real-time system metrics. The analysis reveals that combining complementary detection architectures such as YOLOv8 and RF-DETR Nano under dynamic switching logic represents a promising direction for balancing detection accuracy and inference latency in real-world assistive applications. This survey aims to consolidate current knowledge, highlight technological trends, and guide future research toward more adaptive, efficient, and inclusive assistive vision solutions.***

Index Terms: ***Assistive Technology, Visually Impaired, Object Detection, YOLOv8, RF-DETR Nano, Optical Character Recognition, Text-to-Speech, Speech Recognition, Edge Computing, Hybrid AI, Real-Time Systems, Smart Glasses, Google Coral Edge TPU, Vuzix Blade, Survey***

I. INTRODUCTION

Visual impairment affects hundreds of millions of individuals worldwide, fundamentally restricting their capacity to perform everyday tasks such as navigating unfamiliar environments, identifying objects in their vicinity, and reading textual infor-mation encountered in public spaces. According to the World Health Organization, over 2.2 billion people globally experience some form of vision impairment [9], [10], highlighting the urgent need for accessible and effective assistive technologies. In India alone, approximately 15 million people are blind and 62 million are visually impaired [9], underscoring the scale of the societal challenge.

The proliferation of affordable embedded computing hardware—including single-board computers such as the Rasp-berry Pi—combined with the rapid maturation of deep learning-based computer vision models, has opened new avenues for developing edge-deployable assistive systems. Models from the You Only Look Once (YOLO) family have become the dominant paradigm for real-time object detection in assistive applications, owing to their favorable balance of speed and accuracy. More recently, transformer-based detectors such as RF-DETR Nano have emerged as promising alternatives for low-resource inference scenarios. Dedicated AI accelerator hardware, such as the Google Coral Edge TPU [9], and commercial augmented reality platforms such as the Vuzix Blade 2 [10], have further broadened the hardware ecosystem available for assistive deployments beyond the Raspberry Pi.

However, a persistent challenge in this domain is the fundamental tension between detection accuracy and inference speed. High-accuracy models often demand substantial computational resources and introduce latency that is incompatible with real-time operation, while lightweight models may sacrifice detection quality in complex scenes. A review of the existing literature reveals that most assistive systems adopt a single-model strategy, which precludes dynamic adaptation to varying environmental complexity or computational load. This survey examines ten recent publications in the field of AI-based assistive visual perception, covering wearable smart glasses, edge-deployed detection systems, and multimodal assistive pipelines. The reviewed works are analyzed for their detection strategies, multimodal features, hardware platforms, and identified limitations.

©IJRASET: All Rights are Reserved

Based on this analysis, the survey identifies a research gap in the adaptive, context-aware selection of detection models and discusses the potential of combining complementary architectures such as YOLOv8 and RF-DETR Nano under dynamic switching logic to address the accuracy-latency trade-off.

The remainder of this paper is organized as follows. Sec-tion II presents the literature review. Section III provides an analysis of existing systems and their common technological approaches. Section IV identifies research gaps emerging from the comparative analysis. Section V discusses a conceptual hybrid detection approach derived from the surveyed literature. Section VI outlines research objectives informed by the survey findings. Section VII surveys the general methodologies found in the reviewed literature. Section VIII summarizes commonly used tools and platforms. Section IX concludes the survey.

II. LITERATURE REVIEW

A. *Overview of Reviewed Works*

A substantial body of research has been directed toward leveraging computer vision and artificial intelligence for the benefit of visually impaired individuals. The ten works reviewed in this survey collectively span the period 2024–2026 and represent a variety of architectural approaches, hardware platforms, and evaluation methodologies. The reviewed systems range from wearable smart glasses with multimodal pipelines to edge-deployed detection systems and conceptual assistive frameworks. A common thread across all reviewed works is the integration of object detection as the core visual understanding module, supplemented in many cases by OCR, depth sensing, and audio feedback mechanisms. Newer works additionally explore dedicated AI accelerators (Google Coral Edge TPU [9]) and commercial augmented reality wearables (Vuzix Blade 2 [10]), extending the hardware diversity of the field. A structured comparison of all ten works across method, contribution, application area, and limitation is provided in Table I. The following subsections summarize each reviewed work individually.

B. *Individual Paper Reviews*

Moram et al. [1] presented a multifunctional assistive smart glasses system for visually impaired users that integrates object detection using YOLO-based models, Optical Character Recog-nition, and Text-to-Speech audio feedback into a wearable form factor. The work establishes a foundational multimodal pipeline combining visual detection with audio output. Their qualitative evaluation confirms the practical utility of combining detection and audio feedback in a single assistive platform, though the absence of quantitative performance benchmarks limits comparative assessment. Ruparelia et al. [2] proposed an integrated assistive system that combines YOLO-based object detection across multiple versions (YOLOv5, YOLOv7, and YOLOv8), depth estimation using the MiDaS model, and OCR for text recognition. The system achieved a mean average precision (mAP) of 92.4% and a System Usability Scale (SUS) score of 84.6, demonstrating both high detection accuracy and strong user acceptance. This work highlights the advantage of multi-model evaluation and the significance of integrating depth sensing for spatial awareness. Falcone et al. [3] introduced Eye-Assist, a real-time wearable visual assistance system deployed on mobile smart glasses with depth-sensing capability. The system employs YOLOv8 in TensorFlow Lite (TFLite) format alongside an Intel RealSense depth camera, enabling efficient inference suitable for mobile deployment. The work demonstrates that lightweight versions of YOLOv8 can be successfully deployed on resource-constrained wearable hardware while maintaining functional real-time performance, though explicit performance metrics were not reported. Bailke et al. [4] proposed VisionGuardian, a real-time multi-task detection system that extends standard object detection by incorporating face detection using Multi-task Cascaded Convolutional Networks (MTCNN) and emotion recognition using CNN classifiers. The system employs the Single Shot MultiBox Detector (SSD) as the primary detection backbone. This work illustrates the potential for enriching assistive systems with social interaction features such as facial analysis and emotional state recognition.

Kharat et al. [5] presented AI-powered smart glasses specif-ically designed for outdoor navigation. The system employs YOLOv9 for object detection and integrates ultrasonic sensors and GPS modules for spatial guidance. This work highlights the extension of camera-based detection systems with sensor fusion for reliable outdoor navigation, though it remains narrowly targeted at outdoor environments without adaptive model selection. Noor et al. [6] investigated real-time indoor object detection using YOLOv11 deployed on a Raspberry Pi 4. The system achieved approximately 99% detection accuracy for indoor environments, demonstrating the capability of next-generation YOLO architectures on edge hardware. This result provides a benchmark for detection accuracy achievable on the Raspberry Pi platform. Badawi et al. [7] proposed a smart bionic vision assistive system covering a broad range of features including object detection, face recognition, OCR-based text reading, and color identification using CNN-based models. The conceptual breadth of this work demonstrates the wide variety of visual assistance capabilities that can be integrated into a single system, though it lacks empirical evaluation.

Varghese et al. [8] presented an early real-time object detection and navigation system using the Single Shot MultiBox Detector (SSD) model for guidance in unfamiliar environments. While the SSD-based approach represents an earlier generation of detection technology, the work establishes important ground-work for navigation-oriented assistive systems and motivates the subsequent adoption of more advanced detection architectures. Sharma et al. [9] introduced a conceptual design for AI-powered smart glasses targeting blind and visually impaired individuals, with a strong emphasis on real-world wearability, user dignity, and aesthetic non-intrusiveness. The system employs a hybrid dual-processor architecture: a Google Coral Edge TPU embedded discreetly within the glasses frame for time-critical inference tasks such as obstacle detection (live camera feed analysis), traffic-light recognition using color and shape analysis, and frame-by-frame object classification; and a belt-mounted Raspberry Pi Zero W that handles all non-time-critical higher-level tasks including voice command processing, GPS navigation via the Google Maps API, and wireless communication. A micro wide-angle camera embedded at the lens edge captures a 180-degree field of view. Audio feedback is delivered through bone-conduction speakers that allow simultaneous awareness of environmental sounds. The proposed mathematical decision model governs movement safety through a conjunction of three conditions: the minimum detected obstacle distance exceeding a safe threshold, object classification confidence falling below the obstacle-probability threshold of 0.5, and a confirmed green traffic-light state, formalized as: 16 obstacle categories was constructed from videos recorded under diverse lighting and weather conditions. The comparative evaluation demonstrates that YOLOv8-M achieves the highest accuracy (precision: 0.90, recall: 0.83, mAP50: 0.87) but at computational cost unsuitable for the Vuzix Blade 2's quad-core ARM Cortex-A55 CPU. YOLOv8-S emerges as the optimal deployment model, achieving precision 0.88 and recall 0.81 with inference times of 18–21 ms per frame and overall mAP at 0.5 IoU of 0.877. The peak F1 score across all 16 classes is 0.85 at a confidence threshold of 0.407. The two primary system features are "Walk Safe," which alerts users to any detected obstacle within one meter, and an OCR mode that reads printed text from diverse surfaces (including reflective and low-contrast) and converts it to speech. End-to-end latency from camera capture to speech output is 250–500 ms for object detection and under 500 ms when OCR is triggered. The work identifies three persistent gaps in assistive technology literature: fragmentation between navigation and OCR functionalities, hardware dependencies that restrict real-world deployment, and computational constraints that compromise real-time per-formance [10]. Identified limitations include a short battery life of 2–3 hours for the Vuzix Blade 2, reduced detection performance under extreme lighting, dependency on internet connectivity for Azure OCR, and the absence of formal user trials with visually impaired participants.

C. *Collective Summary*

Across the ten reviewed works, several consistent patterns emerge. Object detection using YOLO-family models domi-nates the literature, with versions ranging from YOLOv5 to YOLOv11. Wearable form factors, particularly smart glasses, $\text{Move}_{\text{Safe}} = (\min(D_i) > D_{\text{threshold}}) \wedge (P(O_i) < 0.5) \wedge (T(t) = 1)$ are the most common deployment platform. Multimodal

where D_i is the distance to obstacle O_i, $P(O_i)$ is the probability of the object being an obstacle, and $T(t) = 1$ denotes a green signal [9]. Benchmarking against ultrasonic canes and mobile navigation applications shows that the proposed glasses achieve 90% obstacle detection accuracy in daylight at sub-100 ms latency, compared to 80% accuracy and 250–500 ms response times for ultrasonic canes. Traffic-light classification confidence exceeds 85%, and the device weighs under 60 g—substantially lighter than conventional cane-based assistive devices (180–400 g). The primary limitations are reduced performance in low-light and night-time scenarios, challenges in NLP processing under ambient noise, and the absence of an empirically tested prototype validated with real users. Kumari and Hammady [10] presented a novel wearable assistive system built on the Vuzix Blade 2 augmented reality smart glasses platform, targeting navigation safety for visually impaired individuals on university campuses. The system integrates real-time object detection using three YOLOv8 model variants (YOLOv8-N, YOLOv8-S, and YOLOv8-M), online object tracking via BoT-SORT, geometric distance estimation, OCR via Azure AI Vision, and text-to-speech via Microsoft Cognitive Speech Services into a unified pipeline. A custom dataset of 15,951 annotated campus images across integration—combining detection with OCR, TTS, and depth sensing—is increasingly prevalent. Edge deployment on af-fordable hardware such as the Raspberry Pi is a shared goal. The two most recent works broaden the hardware landscape: Sharma et al. [9] introduce the Google Coral Edge TPU as an on-glasses AI accelerator enabling sub-100 ms inference, and Kumari and Hammady [10] validate the Vuzix Blade 2 as a commercially viable AR wearable platform achieving mAP 0.877 with sub-500 ms end-to-end latency. These observations are evident from the method and application columns of Table I. However, a critical observation is that no reviewed work employs an adaptive, context-aware model-switching strategy that dynamically selects the most appropriate detector based on real-time system conditions, as reflected in the limitation column of Table I. This gap represents a clear direction for future research in hybrid assistive detection systems, as discussed further in Section IV.

D. Comparative Analysis Table

Table I provides a structured comparison of all ten reviewed works across four dimensions: the primary method or model used, the key contribution, the application area, and the identified limitation.

TABLE I
COMPARATIVE SUMMARY OF REVIEWED WORKS IN ASSISTIVE VISUAL PERCEPTION

Paper / Author	Method / Model	Key Contribution	Application Area	Limitation
Moram et al. [1] (2025)	YOLO, OCR, TTS	Multimodal wearable glasses integrat-ing detection, text recognition, and audio feedback.	Assistive wearables	Qualitative only; no quanti-tative benchmarks.
Ruparelia et al. [2] (2025)	YOLOv5/v7/v8, MiDaS, OCR	Multi-model comparison with depth estimation; mAP 92.4%, SUS 84.6.	Multi-sensor detec-tion	High compute cost; limited edge analysis.
Falcone et al. [3] (2025)	YOLOv8 (TFLite), Re-alSense	Lightweight wearable system with depth sensing for mobile deployment.	Mobile smart glasses	Metrics not reported; scala-bility not analyzed.
Bailke et al. [4] (2025)	SSD, MTCNN, CNN	Multi-task system with object, face, and emotion detection.	Social interaction	SSD less accurate than YOLO; no adaptive switch-ing.
Kharat et al. [5] (2026)	YOLOv9, Ultrasoni c, GPS	Outdoor navigation glasses with de-tection and GPS guidance.	Outdoor navigation	Indoor environments not ad-dressed; no model selec-tion.
Noor et al. [6] (2025)	YOLOv11, Raspberry Pi 4	High-accuracy indoor detection (~99%) on edge hardware.	Indoor edge detection	Indoor only; no speech or OCR integration.
Badawi et al. [7] (2024)	CNN, OCR	Broad-feature system: detection, face, OCR, color recognition.	Multi-feature assis-tive	Conceptual; no empirical evaluation.
Varghese et al. [8] (2024)	SSD model	Early navigation system using SSD for real-time guidance.	Navigation assistance	SSD outperformed by mod-ern models; no speech or OCR.
Sharma et al. [9] (2025)	Coral Edge TPU, RPi Zero W, bone-conduction TTS, GPS	Dual-processor wearable glasses; 90% obstacle accuracy <100 ms; traffic-light recognition >85%; math-ematical safety decision model; de-vice weight <60 g.	Outdoor navigation, wearable design	Low-light performance re-duced; no real-user pro-totype tested; NLP chal-lenged by ambient noise.
Kumari & Hammady [10] (2026)	YOLOv8-N/S/M, BoT-SORT, Azure OCR, Vuzix Blade 2	Campus AR glasses with mAP 0.877; YOLOv8-S optimal (precision 0.88, recall 0.81, 18–21 ms); sub-500 ms end-to-end latency; OCR on reflective/low-contrast surfaces.	University campus navigation, AR wearable	Battery life 2–3 h; no for-mal user trials; cloud OCR requires internet; limited to predefined dataset.

III. ANALYSIS OF EXISTING APPROACHES

A. Technologies Commonly Employed

The literature reveals a well-established set of technologies that underpin assistive visual perception systems, summarized in Table II. YOLO-based object detection models—spanning YOLOv5 through YOLOv11—have emerged as the dominant detection paradigm, owing to their anchor-free or anchor-based architectures that balance detection speed and accuracy. OCR engines such as Tesseract, EasyOCR, and cloud-based Azure AI Vision are widely integrated for text recognition, enabling users to access printed information in the environment [10]. Depth sensing via stereo cameras or LiDAR provides spatial awareness and obstacle distance estimation. Audio feedback through TTS engines and voice command input via STT modules complete the multimodal interaction layer. OpenCV serves as the universal computer vision library across all reviewed platforms [1], [2], [6]. Beyond the Raspberry Pi, the Google Coral Edge TPU [9] and the Vuzix Blade 2 AR platform [10] represent an emerging class of purpose-built or commercial wearable hardware that offer improved performance-per-watt for on-device inference, with the Coral TPU achieving sub-100 ms latency and the Vuzix Blade 2 supporting inference at 18–21 ms per frame via its quad-core ARM CPU.

B. Common Processing Approaches

The typical processing pipeline observed across the reviewed literature consists of the following stages. Image acquisition is performed through a camera mounted on the assistive device. Preprocessing operations—including resizing, normalization, and noise reduction—prepare frames for model inference. A detection model then identifies objects of interest, producing bounding boxes, class labels, and confidence scores. Where applicable, OCR processes the same or a parallel image stream to extract textual content. Depth sensors supplement detection outputs with distance information. The combined outputs are synthesized into audio descriptions delivered to the user via TTS [3]–[5]. In the system of Sharma et al. [9], a formal decision function determines safe movement by jointly evaluating minimum obstacle distance, classification confidence, and traffic-light state. In the system of Kumari and Hammady [10], object tracking via BoT-SORT is applied after detection to maintain object identities across frames, and distance estimation is derived from bounding-box pixel width and known object dimensions using a focal-length formula, enabling safety-zone classification into safe, caution, and danger tiers. Data augmentation strategies including rotation, shear, brightness variation, saturation, blur, and noise injection [10] are employed to improve model robustness under diverse real- world lighting and weather conditions.

C. Identified Limitations Across Literature

A cross-cutting analysis of the reviewed works, as docu-mented in the limitation column of Table I, reveals several recurring limitations. First, all reviewed systems rely on a single detection model, which creates an inflexible pipeline that cannot adapt to varying scene complexity or available computational resources. Second, single-model architectures consistently demonstrate reduced performance when detecting small or partially occluded objects in cluttered scenes [2], [8], [10]. Third, no reviewed system incorporates context-aware model selection—the detection strategy is fixed at design time and does not respond to real-time operating conditions such as CPU load, frame rate, or image resolution. Fourth, the inherent trade-off between detection speed and accuracy in single-model systems cannot be resolved without hardware upgrades, as confirmed by the YOLOv8-N/S/M comparison in which YOLOv8-M achieves higher accuracy than YOLOv8-S but is unsuitable for wearable deployment [10]. Fifth, systems designed for a specific environment—indoor or outdoor—do not generalize effectively to the other [5], [6]. Sixth, battery life remains a persistent constraint: the Coral-based system of Sharma et al. [9] operates for approximately 5 hours, while the Vuzix Blade 2 system of Kumari and Hammady [10] is limited to 2–3 hours. Seventh, the absence of formal user studies with visually impaired participants across most reviewed works [9], [10] limits the ecological validity of the reported performance figures.

IV. RESEARCH GAPS

The comparative analysis of Table I reveals a consistent and critical research gap: *no reviewed work incorporates adaptive, context-aware model selection that dynamically responds to real-time computational and environmental conditions*. All reviewed systems fix a single detection model at design time, precluding any runtime optimization of the accuracy-latency trade-off.

Beyond the core gap of adaptive model switching, the survey identifies several additional gaps. First, the majority of systems are evaluated in controlled or single-environment settings: indoor-only [6] or outdoor-only [5], [9], with few works demonstrating seamless cross-environment generalization. The campus-focused dataset of Kumari and Hammady [10], while more varied, is still limited to predefined university obstacle categories. Second, battery life is universally constrained on wearable platforms, with none of the reviewed works proposing adaptive power management strategies that trade off inference quality against energy consumption.

Third, with the exception of Ruparelia et al. [2] who report a SUS score of 84.6, formal user studies are absent across the reviewed works [9], [10], leaving usability, cognitive load, and long-term adoption behavior untested. Fourth, while Kumari and Hammady [10] demonstrate OCR integration with Azure AI Vision achieving high text recognition accuracy on reflective and low-contrast surfaces, the dependence on cloud connectivity for OCR tasks introduces latency variability and fails in offline scenarios, motivating on-device OCR pipelines. Fifth, the dual-processor architecture of Sharma et al. [9] demonstrates a promising direction for distributing inference load between an on-glasses AI chip and a belt-mounted processor, but this architecture has not been combined with adaptive model switching to further reduce latency during high-resource conditions.

V. DISCUSSION: TOWARD HYBRID DETECTION IN ASSISTIVE SYSTEMS

The gaps identified in the reviewed literature collectively suggest that hybrid detection architectures—combining a high-accuracy model with a low-latency model under adaptive switching logic—represent a promising direction for future research in real-time assistive visual perception.

YOLOv8 emerges from the literature as the most suitable candidate for high-accuracy detection in complex and detail-rich scenes. Its anchor-free detection head and decoupled architecture enable precise localization across a diverse range of object categories, and it has been validated on both server-grade and edge hardware in multiple reviewed works [2], [3], [10]. The systematic comparison of YOLOv8-N, YOLOv8-S, and YOLOv8-M by Kumari and Hammady [10] provides direct quantitative evidence that YOLOv8-S is optimal for resource-constrained wearable platforms, achieving precision 0.88 and recall 0.81 at 18–21 ms per frame, while YOLOv8-M (precision: 0.90) is computationally prohibitive for real-time wearable deployment.

RF-DETR Nano represents an underexplored direction in the assistive systems literature. As a compact detection transformer optimized for low-resource inference, it is well-positioned to complement YOLOv8 in scenarios where computational load is elevated or response latency is critical. Its transformer architecture offers improved handling of occluded and small objects compared to anchor-based detectors.

Adaptive Model Switching is the key capability absent from all reviewed works, as confirmed by the limitation column of Table I. A controller that monitors real-time metrics—image resolution, frames per second, and CPU utilization—and dynamically selects between detection models would resolve the accuracy-latency trade-off that all single-model systems face. Such a mechanism could activate YOLOv8 under favorable resource conditions and fall back to RF-DETR Nano when system load is elevated, yielding consistent performance across varying environments.

The dual-processor concept demonstrated by Sharma et al. [9]—separating time-critical on-glass inference from higher-level navigation tasks on the belt-mounted processor—provides an architectural template that could host such adaptive switching logic, while the systematic model benchmarking of Kumari and Hammady [10] provides empirical baselines against which future adaptive systems should be evaluated.

The integration of OCR for text recognition, STT for voice input, and TTS for audio feedback is consistently validated across the reviewed literature as essential for a complete assistive experience [1], [2], [10]. Figure 1 illustrates the general structure of such a multimodal hybrid pipeline, synthesized from the common architectural patterns observed across the reviewed works.

VI. RESEARCH OBJECTIVES

Based on the gaps identified in the literature and the discussion in Section V, the following research objectives are articulated to guide future work in hybrid assistive visual perception.

The first objective is Real-Time Environmental Perception and Object Recognition. Future work should investigate the design of hybrid detection pipelines capable of identifying objects with high accuracy and low latency, adapting dy-namically to scene complexity and available computational resources. The benchmarking of YOLOv8 variants by Kumari and Hammady [10] provides a quantitative baseline (mAP 0.877 at YOLOv8-S, 18–21 ms per frame) that future adaptive systems should surpass or match while reducing latency variability.

The second objective is Comprehensive Text Accessibility through OCR. Assistive systems should reliably recognize and extract printed text from diverse real-world environments—including signage, product labels, and informational placards—and present it to the user through audio feedback. Cloud-based OCR pipelines such as Azure AI Vision [10] demonstrate high accuracy on reflective and low-contrast surfaces but require internet connectivity; future work should investigate on-device alternatives to ensure robust offline operation.

©IJRASET: All Rights are Reserved

The third objective is Enhanced Social Interaction via Facial Analysis. Research should explore the integration of face detection and emotion recognition to enrich the social interaction capability of assistive systems, enabling users to identify individuals and interpret emotional context, as pioneered by VisionGuardian [4].

The fourth objective is Integrated Navigation and Spatial Wayfinding. Assistive frameworks should provide contextual navigational guidance by conveying the location and relative position of detected objects and obstacles, supporting safe and independent user movement. The GPS-integrated and traffic-light-aware approach of Sharma et al. [9] and the proximity-zone classification of Kumari and Hammady [10] represent two complementary strategies toward this goal. The fifth objective is Optimized Edge Performance and Accessibility. Future systems in this domain should be designed to operate on low-cost, portable edge hardware without dependence on cloud-based inference, ensuring practical de-ployability for users across diverse socioeconomic contexts. The Coral Edge TPU architecture [9] and the on-device TFLite deployment of Eye-Assist [3] represent concrete steps in this direction, while the Vuzix Blade 2 system [10] demonstrates the viability of commercial AR wearables as an alternative deployment platform.

VII. GENERAL METHODOLOGIES USED IN LITERATURE

A. *Overview of Common Pipelines*

The reviewed literature converges on a modular pipeline architecture for real-time assistive visual perception. While specific implementations differ across works, the general structure follows a consistent sequence of stages: input ac-quisition, image preprocessing, object detection, supplementary recognition (OCR or depth), decision aggregation, and audio output generation. This section surveys each stage as it appears across the reviewed works.

B. *Input Acquisition*

Visual input is universally captured through a camera module—either a standard webcam, a dedicated depth camera such as Intel RealSense [3], a Pi Camera Module integrated into a wearable device [6], a micro wide-angle camera embedded discreetly in the glasses frame [9], or the 8 MP front-facing camera of Vuzix Blade 2 AR glasses capable of 1080p video capture [10]. In systems that support voice interaction, a microphone captures user commands concurrently with the visual stream; Sharma et al. [9] integrate a noise-cancelling microphone in the temple arm specifically to withstand outdoor ambient noise.

C. *Image Preprocessing*

All reviewed systems apply a preprocessing stage before model inference. Common operations include resizing frames to match the model's expected input dimensions, applying noise reduction to mitigate sensor or lighting artifacts, and nor-malizing pixel intensities to ensure consistent input scaling [2], [6]. Kumari and Hammady [10] additionally apply contrast stretching and data augmentation (rotation, shear, brightness, saturation, blur, noise) to improve model generalization across diverse lighting and weather conditions encountered on univer-sity campuses.

D. *Object Detection Strategies*

Object detection is the core visual processing stage across all reviewed works. YOLO-based detectors dominate the literature, with YOLOv5, YOLOv7, YOLOv8, YOLOv9, and YOLOv11 each appearing in one or more reviewed systems [2], [3], [5], [6], [10], as catalogued in the method column of Table I. SSD-based detectors appear in earlier work [4], [8]. The Google Coral Edge TPU in Sharma et al. [9] runs TPU-compiled classification models enabling sub-100 ms on-glass inference without cloud dependency, while the Vuzix Blade 2 system of Kumari and Hammady [10] executes YOLOv8-S via TensorFlow Lite with NNAPI hardware acceleration, achieving 18–21 ms per frame. Kumari and Hammady [10] provide the only head-to-head comparison of multiple YOLO variants (N, S, M) on the same dataset and hardware, establishing YOLOv8-S as the optimal balance point for commercial wearable deployment. All reviewed systems employ a single, fixed detection model—no work combines multiple detectors under runtime switching logic. Following detection, Kumari and Hammady [10] apply BoT-SORT tracking (Kalman filtering combined with the Hungarian algorithm) to maintain object identities across video frames, enabling continuous proximity monitoring.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

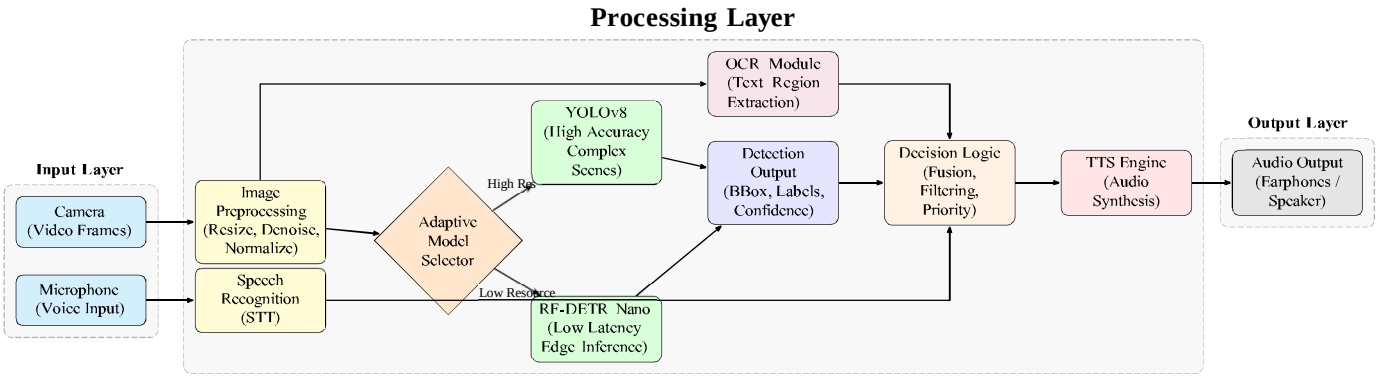


Fig. 1. General workflow of a hybrid multimodal assistive visual perception pipeline, synthesized from architectural patterns observed across the reviewed literature [9], [10]. An adaptive model selector dynamically routes inference to either YOLOv8 or RF-DETR Nano based on real-time system metrics. Detection outputs, OCR results, and speech commands are fused by the decision logic module before audio synthesis and delivery to the user. The dual-processor pattern observed in Sharma et al. [9]—on-device AI chip plus external processor—motivates the separation of time-critical inference from higher-level navigation logic in this architecture.

A. Text Recognition and Depth Estimation

Several reviewed systems augment object detection with OCR for text recognition, enabling users to access printed content in the environment [1], [2], [7], [10]. OCR is applied either to the full frame or to detected regions of interest. Azure AI Vision provides cloud-based OCR in the Vuzix Blade 2 system [10], demonstrating high accuracy on reflec-tive, low-contrast, and uneven surfaces with recognized text converted to speech via Microsoft Cognitive Speech Services. Depth estimation, using either MiDaS [2] or dedicated depth cameras [3], provides distance information that complements bounding box detection for obstacle avoidance and navigation guidance. Kumari and Hammady [10] adopt a lightweight geometric distance estimation approach using the formula $d = (f \times W_{\text{real}})/W_{\text{pixel}}$ (focal length multiplied by real object width, divided by pixel width), providing safety-zone classification without requiring dedicated depth hardware.

B. Speech Interaction

Voice-based interaction is present in a subset of the reviewed systems. STT modules convert user voice commands into text, enabling natural interaction with the assistive pipeline [1], [9]. TTS engines convert system outputs—detection descriptions, OCR results—into audio feedback delivered through earphones or speakers. Sharma et al. [9] use bone-conduction speakers to deliver audio guidance while preserving the user's awareness of environmental sounds, which is particularly important in outdoor navigation. Kumari and Hammady [10] leverage Microsoft Cognitive Speech Services for natural-sounding text-to-speech output, though they note reduced clarity in high ambient-noise outdoor settings and propose bone-conduction output as a future enhancement. The combination of STT and TTS enables hands-free, eyes-free operation essential for effective assistive use.

C. Decision Aggregation and Output

The final stage across all reviewed pipelines involves aggregating the outputs from detection, OCR, and speech modules into a coherent, priority-ordered audio description. Priority-based filtering is applied in more sophisticated systems to prevent information overload [5]. Sharma et al. [9] introduce a formal mathematical decision model that gates navigation commands on a three-way conjunction: minimum obstacle distance exceeding a safe threshold, object classification confidence below the obstacle-probability threshold (0.5), and a confirmed green traffic-light state. Kumari and Hammady [10] implement a "Walk Safe" alert triggered whenever any tracked object enters the one-meter proximity zone, with safety-zone feedback (safe, caution, danger) delivered continuously during navigation. Audio output is delivered using TTS engines such as pyttsx3, gTTS, or Microsoft Cognitive Speech Services.

VIII. COMMONLY USED TOOLS AND PLATFORMS

Table II summarizes the hardware components and software tools most commonly referenced across the reviewed literature, expanded to include the Google Coral Edge TPU and Vuzix Blade 2 platforms validated in the two most recent works [9], [10].

IX. CONCLUSION

This survey has reviewed ten recent works in the domain of AI-based assistive visual perception systems for visually impaired users, spanning the period 2024–2026. The reviewed works collectively demonstrate significant progress in the application of deep learning, edge computing, and multimodal interaction to assistive technology. Several important trends are observed from the literature.

First, YOLO-family models—from YOLOv5 to YOLOv11—have become the undisputed standard for real-time object detection in assistive systems, offering a scalable trade-off between accuracy and speed. The systematic three-way com-parison of YOLOv8 variants by Kumari and Hammady [10] confirms that YOLOv8-S is the optimal deployment choice for resource-constrained wearable hardware, achieving mAP 0.877 at an inference time of 18–21 ms per frame on a

TABLE II Summary of Commonly Used Hardware and Software in Reviewed Literature

Category	Component / Tool	Purpose
Hardware Platforms		
Processing Unit	Raspberry Pi 4 Model B	Edge inference host
Processing Unit	Raspberry Pi Zero W	Navigation & voice processing [9]
AI Accelerator	Google Coral Edge TPU	On-glass low-latency AI inference [9]
AR Wearable	Vuzix Blade 2	Commercial AR smart glasses [10]
Visual Input	Pi Camera / RealSense	Frame capture / depth
Audio Input	Noise-cancelling Mi-crophone	Voice command cap-ture [9]
Audio Output	Earphones / Bone-conduction Speaker	TTS delivery [9]
Power Supply	Portable Power Bank	Mobile operation
Storage	SD Card (16–32 GB)	OS and model storage
Software Tools		
Language	Python 3.x	Primary development
Vision Library	OpenCV	Frame processing
Detection Model	YOLOv8 (Ultralytics)	High-accuracy detec-tion
Detection Model	RF-DETR Nano	Low-latency detection
OCR Engine	Tesseract / EasyOCR	Local text recognition
OCR Service	Azure AI Vision	Cloud OCR [10]
TTS Service	Microsoft Cognitive Speech	Cloud TTS [10]
STT Library	SpeechRecognition (Py)	Voice command input
TTS Library	pyttsx3 / gTTS	Local audio feedback
Navigation API	Google Maps API	Turn-by-turn GPS guid-ance [9]
Object Tracker	BoT-SORT	Multi-frame object tracking [10]
Dev Framework	Unity (Android)	AR app development [10]

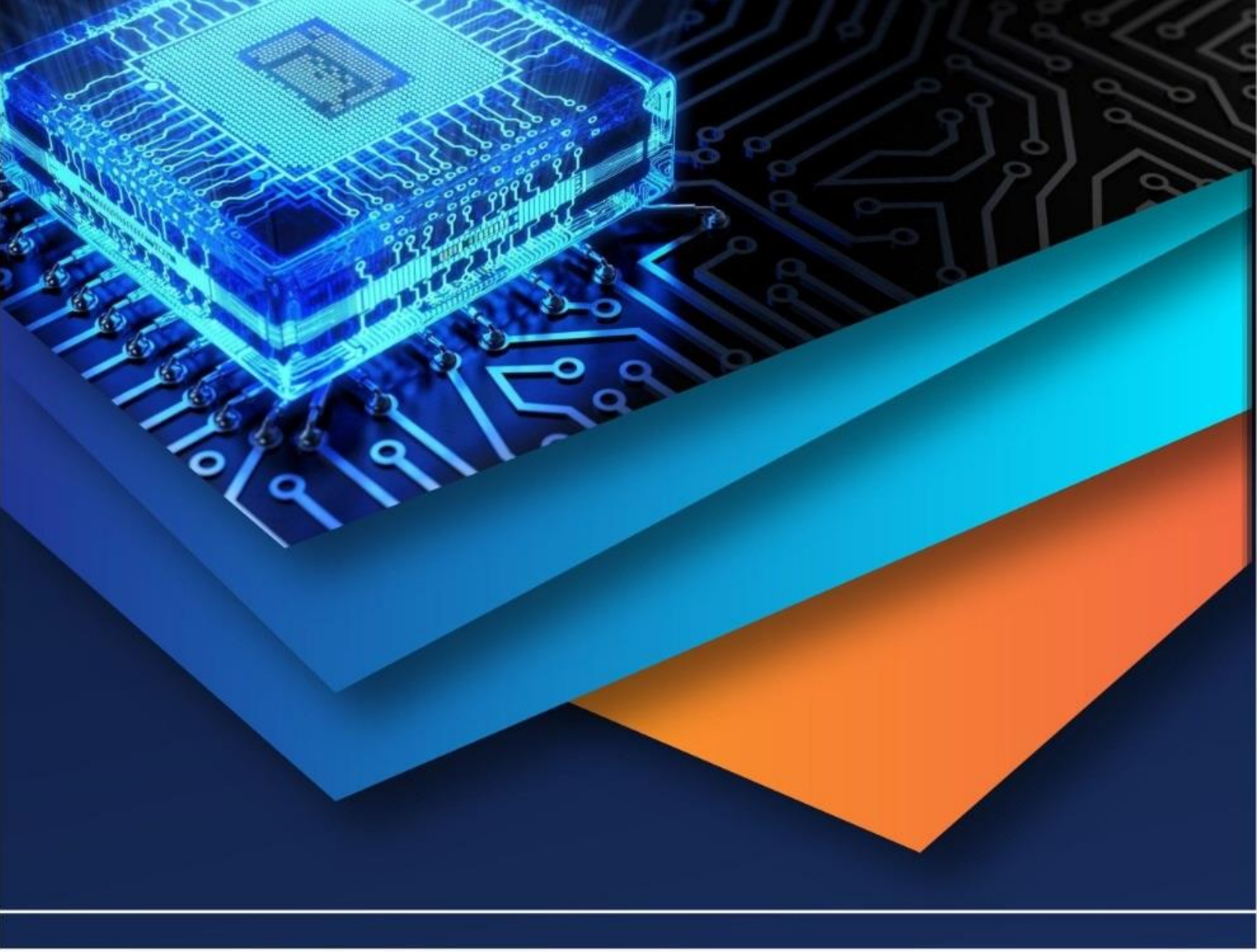
quad-core ARM platform. Second, multimodal integration—combining detection with OCR, TTS, STT, and depth sensing—is increasingly recognized as essential for delivering a complete and usable assistive experience [1], [2], [10]. Third, edge deployment on affordable single-board computers such as the Raspberry Pi has been validated as a technically feasible and cost-effective approach, with detection accuracy approaching 99% on such hardware [6]. Fourth, dedicated AI accelerators such as the Google Coral Edge TPU [9] and commercial AR wearables such as the Vuzix Blade 2 [10] are emerging as viable hardware platforms that extend beyond the Raspberry Pi ecosystem, enabling sub-100 ms inference and richer user interfaces. Fifth, the dual-processor architecture of Sharma et al. [9]—separating on-glasses time-critical inference from belt-mounted higher-level processing—provides a practical and aesthetically non-intrusive deployment pattern that achieves 90% obstacle detection accuracy at sub-100 ms latency for a device weighing under 60 g. The hardware and software components underpinning these systems are consolidated in Table II.

©IJRASET: All Rights are Reserved

However, the comparative analysis of Table I reveals a consistent and critical limitation: no reviewed system incorpo-rates adaptive, context-aware model selection that dynamically responds to real-time computational and environmental con-ditions. All reviewed systems fix a single detection model at design time, precluding any runtime optimization of the accuracy-latency trade-off. Addressing this gap through hybrid detection architectures—such as a dynamic YOLOv8 / RF-DETR Nano switching controller inspired by the dual-processor pattern of Sharma et al. [9] and benchmarked against the empirical baselines of Kumari and Hammady [10]—represents the most significant open research direction identified by this survey.

REFERENCES

[1] V. Moram, S. Zahruddin, and S. Kumar, "Multifunctional assistive smart glasses for visually impaired," SN Computer Science, vol. 6, no. 2, pp. 1–12, Mar. 2025. DOI: 10.1007/s42979-025-03456-x
[2] K. Ruparelia, P. Parikh, and P. Shah, "Integrated assistive system using YOLO-based detection, depth estimation, and OCR," American Journal of Computer Science and Technology, vol. 8, no. 1, pp. 45–58, Jan. 2025.
[3] D. C. S. Falcone, A. Brown, F. Salim, et al., "Eye-Assist: Real-time visual assistance system for the visually impaired," in Proc. IEEE 16th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 2025, pp. 112–118.
[4] P. A. Bailke, A. Gorave, O. Bhojane, et al., "VisionGuardian: Real-time multi-task detection system for visually impaired users," in Lecture Notes in Networks and Systems, vol. 1012, Springer, Singapore, 2025, pp. 233–245.
[5] R. Kharat, S. Thepade, A. Kothawade, et al., "AI-powered smart glasses for outdoor navigation using computer vision," in Lecture Notes in Networks and Systems, vol. 1085, Springer, Singapore, 2026, pp. 317–328.
[6] A. Noor, H. Almukhalfi, A. Souza, and T. H. Noor, "Real-time indoor object detection using YOLOv11 on Raspberry Pi 4," Computer Modeling in Engineering & Sciences, vol. 142, no. 1, pp. 211–228, 2025. DOI: 10.32604/cmes.2025.058312
[7] M. I. Badawi, A. J. Al-Nagar, R. S. Mansour, et al., "Smart bionic vision assistive system for visually impaired individuals," Biomedical Journal of Scientific & Technical Research, vol. 58, no. 2, pp. 47843–47851, Jun. 2024.
[8] N. Varghese, S. Agrawal, and M. K. Gupta, "Real-time object detection and navigation assistance using SSD for the visually impaired," in Proc. 2024 International Conference on Advances in Computing and Communications (ICACC), Kochi, India, 2024, pp. 1–6.
[9] P. Sharma, A. S. Babu, M. Sadasivan, K. T. A. Robert, M. T. Vasumathi, and A. K. Ashok Kumar, "Smart glasses for the blind: A real-time AI-driven wearable system for autonomous navigation," in Proc. 2025 International Conference on Computing and Communications (COM-PUTINGCON), Pune, India, Sep. 2025, pp. 1–7. DOI: 10.1109/COM-PUTINGCON64838.2025.11379829
[10] P. Kumari and R. Hammady, "Assisting blind people with AI and audio using smart glasses: system design with YOLOv8 variants comparisons," Multimedia Systems, vol. 32, no. 73, Jan. 2026. DOI: 10.1007/s00530-025-02139-z



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)