Multimedia Systems (2026) 32:73
https://doi.org/10.1007/s00530-025-02139-z

REGULAR PAPER

Assisting blind people with AI and audio using smart glasses: system design with YOLOv8 variants comparisons

Priyanka Kumari[1] · Ramy Hammady[2,3]

Received: 29 June 2025 / Accepted: 5 December 2025

Abstract

This paper introduces a novel system design leveraging Vuzix Blade 2 smart glasses to enhance the mobility and independence of visually impaired individuals. The study critically examines existing assistive navigation and object detection technologies, identifying their limitations and gaps. The designed system integrates real-time object detection, distance estimation, and OCR functionalities, providing auditory feedback through a robust and efficient pipeline. The designed application enhances the independence and safety of visually impaired individuals, particularly in navigating university campuses. A dataset comprising 15,951 annotated images from the university campus was used for training and evaluation. A comparative analysis of three YOLOv8 models (YOLOv8-N, YOLOv8-S, and YOLOv8-M) was conducted, balancing accuracy and computational efficiency to optimise system performance. The pipeline also offers a comprehensive framework for developers and researchers to build inclusive systems combining AR, computer vision, and AI. Results show high object detection accuracy (precision: 0.90, recall: 0.83) and reliable distance estimation with a minor error of 0.33 m. Results demonstrate the system's capability to detect obstacles within one meter, provide precise distance estimation, and convert textual information into speech, validating its potential for real-world applications. This study emphasises the significant role of AI-driven solutions in advancing assistive technologies, paving the way for more accessible and inclusive navigation systems. Compared with recent assistive systems such as Smart Cane (He in CCF Trans. Pervasive Comput. Interact. 5:382–395, 2023), OrCam MyEye (Amore in J. Med. Syst. 47:11, 2023), and IrisVision (Gopalakrishnan in Comparison of visual function analysis of people with low vision using three different models of augmented reality devices, 2024), the proposed system demonstrates superior integration of detection, text recognition, and real-time feedback within a lightweight wearable device.

Keywords Assisted technology · Augmented reality · Object detection · OCR · AR application

1 Introduction

Eyesight is fundamental to human interaction with the environment, playing a critical role in navigating and interpreting visual information. However, for the visually impaired, even simple day-to-day activities can become challenging, potentially leading to physical and mental health issues, such as increased risk of accidents, loss of confidence, and social isolation [62]. As of 2020, over 1.1 billion individuals worldwide have experienced some form of vision loss, with projections indicating a significant increase by 2050 [32]. These figures underscore the urgent need for accessible and effective assistive technologies.

Despite advancements in assistive technologies, significant gaps remain in providing visually impaired individuals with intuitive, -time, and comprehensive solutions.

Communicated by Mang Ye

✉ Ramy Hammady
r.hammady@soton.ac.uk

Priyanka Kumari
priyankagoraikumari@gmail.com

[1] Harvy Nash, London, UK

[2] Design Department, Winchester School of Art, University of Southampton, Winchester, UK

[3] Faculty of Applied Arts, Helwan University, Giza, Egypt

Published online: 20 January 2026

Many existing tools focus on isolated functionalities, such as obstacle detection or text recognition, without seamlessly integrating these features into a cohesive system. For example, NavCog provides internal navigation support, for instance at Pittsburgh Airport, while Moovit focuses mainly on public transport by offering route guidance, notifications, and travel warnings. Devices such as Ultracane and Miniguide, which use sonar-based obstacle detection, can assist users in navigating their surroundings but are still limited in object detection and -time responsiveness. They are unable to track moving obstacles, which can cause confusion during travel. This highlights how focusing on a single feature does not make users fully independent in their navigation [47]. Similarly, OCR tools such as "Seeing AI", which drew inspiration from Microsoft Kinect, rely on manual channel switching for detecting & reading various objects & texts, limiting their usability in dynamic environments. NAVI, which used Microsoft Kinect, was very bulky, needed a backpack to carry and had low battery life. Like Ultracane and Miniguide, NAVI, which relied on a depth histogram for object detection, did not capture the dynamic movement of objects[72].

Furthermore, existing object detection systems, including those integrated into smart glasses or mobile applications, often struggle with accuracy in real-time scenarios due to hardware limitations and computational constraints. Traditional models such as the Histogram of Oriented Gradients (HOG) introduced human detection capabilities that supported early navigation systems; however, they lacked diversity in object recognition and were limited to detecting humans in static images [15]. Although recent iterations, such as YOLOv8 and YOLO-NAS, address these challenges by leveraging advanced neural architectures, their implementation in assistive technologies remains limited [65].

Another critical limitation is the lack of user-centric design in these technologies. Many devices require extensive adaptation or training for users, which can be a barrier for visually impaired individuals. For instance, "EyeMusic" and "vOICe" convert visual information into auditory cues but demand significant learning effort, reducing their accessibility [11]. The generated audio representations are often slow and information-dense, resulting in high cognitive load and delayed response time during navigation [43]. Additionally, both lack object recognition, real-time tracking, and semantic understanding, offering only low-level perceptual encoding of shapes and brightness rather than meaningful environmental context [45].

Also, some solutions, like "Aira," rely on remote human agents who perform manual object identification by observing the live video stream transmitted from the user's smartphone or smart glasses. These trained agents verbally describe the user's surroundings, read visible text, and assist in tasks such as navigation or locating objects within the camera's field of view, introducing latency and privacy concerns [44].

Recent publications have highlighted the potential of combining advanced AI models with user-friendly interfaces to overcome these limitations. Although recent iterations such as YOLOv8 and YOLO-NAS leverage advanced neural-architecture search and backbone optimisations [57], their deployment in assistive technologies for visually impaired users remains limited due to constraints such as device computational power, real-world robustness and integration overhead. Recent comparative studies have demonstrated progress toward integrating vision-based assistive systems using deep learning and wearable hardware. For instance, C. He and Saha [27] proposed a Smart Cane using depth cameras for obstacle detection, but lacked OCR or auditory feedback. Gopalakrishnan et al. [22] presented IrisVision, which enhances residual vision but offers no object classification. Amore et al. [4] evaluated OrCam MyEye, highlighting high accuracy in text reading yet a limited field of view (45°) and no real-time navigation. In contrast, the proposed system integrates YOLOv8-based object detection, distance estimation, and Azure-based OCR within a single wearable framework capable of real-time processing on Vuzix Blade 2 smart glasses. This integration addresses key shortcomings of existing solutions—particularly the lack of multimodal fusion and latency-free feedback—thus providing a more holistic assistive experience [1, 2].

Similarly, advances in OCR, such as the use of EAST with recurrent neural networks, enable efficient text recognition even under challenging conditions [68]. However, the integration of these state-of-the-art technologies into a unified, accessible solution for visually impaired users is still an underexplored area. Therefore, this study aims to answer the following question:

RQ: How can a unified assistive system integrate real-time object detection, OCR, and auditory feedback to address the limitations of fragmented solutions in existing assistive technologies for visually impaired individuals?

This research presents an integrated application capable of running computationally intensive models (YOLOv8) efficiently while combining multiple features such as OCR, cloud-based text-to-speech, and real-time object detection and tracking within lightweight smart glasses. The result is a compact, accessible, and user-friendly system that supports visually impaired users across a variety of real-world situations.

Previous assistive devices often struggled with limitations such as the absence of dynamic object tracking and restricted accuracy [19, 53]. While advanced models like YOLO offered higher detection precision, they demanded substantial computational resources, making the systems bulky and difficult to deploy in real time [1, 2]. Moreover, most existing applications were designed for specific contexts, either indoor or outdoor and offered isolated functionalities rather than a comprehensive solution [25, 38].

This research seeks to address the previously mentioned gaps by developing a mobile application that combines real-time object detection with OCR functionality. By leveraging cutting-edge technologies like YOLOv8 and Azure AI Vision, and training on a custom dataset tailored to a university environment, the application aims to provide a seamless, intuitive, and efficient solution for visually impaired users. The main objective of this project is to develop an application for visually impaired individuals to navigate safely in the outdoors of a university campus without assistance. Unlike existing fragmented assistive tools, the proposed system provides an integrated and real-time wearable solution combining object detection, distance estimation, and OCR functionalities, ensuring low latency, enhanced autonomy, and improved usability for visually impaired users.

This application includes two key features: OCR with speech functionality and a safe walk feature that alerts users to potential collisions within one meter. The application has two primary features. One is OCR with speech functionality, as this feature reads out any text of interest to the user. For example, the user can take a picture of a menu in the canteen, and the OCR will extract the text, which the speech function will read aloud. The second feature is "walk safe", which helps the user navigate safely by updating them about potential collisions within one meter. The user can detect streetlights, cars, or any individual passing within one meter.

This design empowers visually impaired individuals to navigate and manage a wide range of everyday situations, such as moving between classrooms, attending conferences, reading text, or independently ordering food in a cafeteria. While the current application has been trained and optimised for use within university premises, its framework can be readily extended to broader environments such as hospitals, workplaces, and other public spaces.

To validate this approach, the remainder of this paper is structured as follows: Sect. 2 critically reviews existing assistive technologies to clarify the research gap; Sect. 3 details our integrated system architecture and the comparative evaluation methodology for YOLOv8 variants; Sect. 4 presents performance results; Sect. 5 discusses implications relative to existing systems; and Sect. 6 concludes with limitations and future directions.

2 Literature review

The advancement of assistive technologies has significantly improved the quality of life for visually impaired individuals. However, limitations in intuitiveness, real-time performance, and the integration of multiple functionalities persist. This literature review explores existing technologies and methodologies relevant to object detection, OCR, and their integration into assistive systems while identifying gaps addressed by this research.

2.1 Assistive technologies for navigation and object detection

Navigation aids have evolved to support visually impaired individuals by detecting obstacles and guiding them through environments. Early systems, such as the Microsoft Kinect-powered NAVI, focused on obstacle detection and recognition, offering mobility enhancements but limited to indoor use due to hardware dependencies [72]. Similarly, systems like Smart Cane and Haptic Radar employed depth cameras and infrared sensors to detect nearby objects, but their reliance on specific hardware constrained scalability and adoption [33]. Recent innovations in object detection have leveraged advancements in deep learning. YOLO (You Only Look Once) models have revolutionised real-time object detection with their speed and accuracy. Earlier versions, such as YOLOv3, introduced multi-scale predictions, while YOLOv4 incorporated enhancements like Mish activation and spatial pyramid pooling [48]. The latest iterations, including YOLOv8, have further optimised detection through anchor-free architectures and efficient backbone networks, enabling faster and more accurate detections in dynamic environments [65]. Despite these advancements, the integration of these models into practical assistive applications remains limited, primarily due to computational constraints and the lack of custom datasets tailored for specific environments. Table 1 provides an overview of assistive technologies developed over the years, as well as their innovations and limitations.

2.2 OCR in assistive applications

OCR technology plays a crucial role in enabling visually impaired individuals to interpret textual information. Early OCR systems, such as Tesseract, relied on handcrafted features and template matching, which limited their performance under varying conditions [54]. Modern OCR approaches, including EAST (Efficient and Accurate Scene Text Detector) and its integration with neural networks like LSTMs, have significantly improved accuracy and

Table 1 Review of assistive technologies

Study	Technology Used	Innovation	Key Features	Applications	Advantages	Disadvantages
[72]	Microsoft Kinect	NAVI	Obstacle detection	Navigation systems	Low latency	Limited to indoor
[47]	Project Tango	SLAM technology	Indoor positioning	Drones	High accuracy	Hardware dependent
[25]	Depth camera	Smart Cane system	Obstacle detection	BVI navigation	Enhanced mobility	Specific hardware
[68]	Project Tango, (Phab 2), haptic actuators	ISANA	Indoor wayfinding, obstacle detection	BVI navigation	Utilises haptic feedback	Hardware dependent
[38]	Tango, Unity	Prototype for indoor environment virtual replica	Capture user's movement, continuously updated virtual replica	Wayfinding, mobility assistance	Real-time environment mapping	Requires game engine integration
[33]	3D image enhancement	Smart Specs	Enhanced 3D perceptions with simplified images emphasising depth	Guidance for VI people	Exploits residual vision	Limited market access
[42]	IR-based system, virtual sound guidance	Haptic Radar	Obstacle avoidance, virtual sound guidance	Pedestrian route guidance	Positive after-test appraisals	Limited area coverage by IR sensors
[11]	3D model, ultrasonic-based motion capture system	Virtual Haptic Radar	Warning vibrations near objects	Tactile-based navigation	Introduces virtual tactile elements	Bulkiness of portable haptic interfaces
[67]	BLE	NavCog smartphone application	Indoor wayfinding	Navigation systems for BVI users	Utilises BLE for precise indoor navigation	Limited to indoor use
[60]	Visual– auditory systems	EyeMusic	Sensory substitution	Visual-to-auditory information conversion	Makes visual information accessible through sound	Requires learning and adaptation
[15]	Visual– auditory systems	The classic vOICe	Sensory substitution	Visual-to-auditory information conversion	Long-standing, well-tested solution	User adaptation required
[52] [37]	Cloud computing, image recognition	Seeing AI, TapTapSee	Verbal image descriptions	Assistive technology for the visually impaired	Provides verbal descriptions of images using remote processing	Dependent on internet connectivity
[26]	Public transport data integration	Moovit	Real-time public transport guidance	Mobility assistance for BVI users	Free, effective, easy-to-use	Limited to areas with public transport data
[21]	GPS, Foursquare's and OpenStreetMap's databases	BlindSquare	Speech-based POI location	Outdoor navigation for BVI users	Designed specifically for BVI users	Depends on external database accuracy
[50]	GPS, motion capture and orientation sensors	Lazzus	Intuitive cues about POI locations	Outdoor navigation for BVI users	Provides verbal information about nearby POIs	Paid application
[50]	GPS, similar to Lazzus	Seeing AI GPS	360° and beam modes for POI information	Outdoor navigation for BVI users	Incorporates pre-journey information	N/A

robustness in detecting and recognising text in complex scenes [69].

Applications such as Seeing AI and TapTapSee utilise cloud-based OCR to provide verbal descriptions of visual content. While these systems offer high accuracy, they are dependent on internet connectivity, making them less effective in offline or remote scenarios [47]. Additionally, the lack of integration between OCR and real-time object detection in these solutions limits their ability to provide a holistic assistive experience. For example, Seeing AI excels in reading text from images but lacks the capability to detect objects in real-time, which diminishes its usability in dynamic environments. Table 2 summarises significant assistive solutions for OCR and navigation, as well as their benefits and limitations.

A critical review of Table 2 reveals persistent limitations in these technologies. Although innovative, solutions like eSight 3 and Eyesynth are prohibitively expensive, limiting accessibility for a broader audience. Moreover, devices such as Oton Glass focus narrowly on dyslexic users, excluding a wider population of visually impaired individuals who require comprehensive features, including proximity

Table 2 Review of industrial assistive technologies

Solutions	Developer	Conceptual Design	Benefits	Drawbacks	Improvements
eSight 3	CNET's	High-res image and video capture	Surgery-free aid	Does not improve vision as it just an aid	Waterproof versions are under development
Oton Glass	Keisuke Shimakage, Japan	Image-to-audio conversion	Symbols to audio conversion, normal looking glasses, supports languages	For only people with reading difficulty and no support for blind people	Can be improved to support blind people also by including proximity sensors.
Aira	Suman Kanuganti	Smart glasses & remote agents	Help users to interpret their surroundings with smart glasses.	Waiting time connected to the Aira agents in order to be able to sense.	To include language translation features.
Eyesynth	Eyesynth, Spain	3D scene-to-audio conversion	Allows blind/ limited sight people to 'feel the space' through sounds. It converts spatial and visual information into audio.	Costly, and it only recognizes objects and directions.	Can use verbal audio for a better feel and navigation services.
Google Glasses	Google Inc.	Hands-free interaction, Voice commands	Can capture images & videos, get directions, send messages, audio calling and real-time translation using word lens app.	Costly, and the glasses are not very helpful for blind people.	Reduce costs to make it more affordable for the consumers.
OrCam MyEye	ORCAM	Daily assistance	Text reading, facial recognition, Standalone operation, no internet required	Limited Field of View (FOV) (45°), no navigation, high cost	Reduce costs to make it more affordable, to be open for customisation and third parties' integrations.

detection and navigation. While Aira offers guided assistance, its dependency on remote human agents introduces latency and privacy concerns, highlighting the need for autonomous systems.

Recent advancements in OCR research have focused on improving robustness under diverse conditions. For instance, models combining EAST with LSTM networks have demonstrated significant potential for end-to-end text recognition, particularly in cluttered or low-light scenarios (Zhang et al., 2020). Additionally, advancements like Google's Vision AI provide scalable and efficient OCR capabilities, but integration into real-time systems remains challenging (Deci AI, 2024). Among the most commercially viable assistive devices for the visually impaired is OrCam MyEye, which integrates a miniature camera with AI-based OCR to read text, recognise faces, and identify products. The device attaches magnetically to eyeglasses and provides real-time audio feedback. Unlike Seeing AI or Aira, which depend heavily on mobile devices or human agents, OrCam MyEye is self-contained and functions offline, addressing critical challenges of latency and privacy[23]. However, despite its autonomy, OrCam's closed system architecture limits customisation and third-party integrations, making it less flexible for research-driven or user-specific enhancements. Additionally, the cost remains prohibitively high for many users [4]. Comparatively, eSight Eyewear uses a combination of high-definition cameras and OLED screens to enhance residual vision, but it is more suitable for users with partial sight rather than total blindness [40]. IrisVision, another contender, relies on a smartphone-based headset and provides magnification and scene interpretation but lacks integration with OCR or robust object detection pipelines [22].

The overarching challenge lies in developing an affordable, real-time OCR solution that seamlessly integrates with object detection for comprehensive assistance. Future systems must bridge this gap by leveraging advanced AI models and optimising for low-resource devices, enabling broader adoption among visually impaired populations. Recent evaluations of wearable assistive technologies highlight a growing shift toward AI-driven, head-mounted systems that enable real-time scene interpretation, object recognition, and reading assistance.

2.3 Advances in object detection

The advent of deep learning marked a paradigm shift in object detection. R-CNN (2014) combined selective search with convolutional neural networks (CNNs) to improve accuracy but faced challenges with computational efficiency [21]. Successive innovations like Fast R-CNN (2015) streamlined the training process for bounding box regressors and detectors, while Faster R-CNN (2015) integrated a Region Proposal Network (RPN) to achieve near real-time performance [5]. Feature Pyramid Networks (FPNs), introduced in 2017, further advanced detection systems by enabling multi-scale feature representation, improving accuracy on benchmarks such as the COCO dataset [21, 39].

One-stage detectors, including YOLO, SSD, and RetinaNet, revolutionised object detection by balancing speed and accuracy. Unlike two-stage methods, these models eliminated region proposal steps, enabling faster inference suitable for real-time applications. YOLO, introduced in 2016, achieved unprecedented speed but faced challenges with localisation accuracy, particularly for smaller objects. Subsequent models, such as SSD, enhanced accuracy with multi-resolution techniques, while RetinaNet introduced focal loss to address the imbalance between easy and hard examples during training [39, 48].

CornerNet and CenterNet adopted keypoint-based detection paradigms, simplifying the detection process by eliminating anchor boxes, which streamlined computational requirements while maintaining competitive accuracy. Recent advancements, including DETR and Deformable DETR, integrated Transformer-based architectures to predict object sets end-to-end without anchor boxes, achieving state-of-the-art results on COCO datasets [10, 66].

2.4 Evolution of the YOLO framework

The YOLO framework has undergone significant development over the years, introducing innovative features and addressing the limitations of its predecessors. YOLOv2 improved upon the original YOLO by incorporating enhancements such as BatchNorm, higher resolution input, and anchor boxes, which increased accuracy and adaptability across various applications [48]. YOLOv3 further advanced the framework by integrating an objectness score into bounding box predictions and generating multi-scale predictions to improve detection performance for smaller objects [13, 58]. YOLOv4 introduced feature aggregation techniques, Mish activation, and augmentation strategies, achieving higher accuracy and speed than its predecessors [9]. With YOLOv5, emphasis was placed on deployment flexibility, providing support for batch processing and seamless conversion into formats like ONNX and CoreML for compatibility with various platforms [36]. Baidu's PP-YOLO and PP-YOLOv2 models enhanced the YOLO framework by integrating techniques such as DropBlock regularisation and Matrix NMS, resulting in superior performance metrics on COCO benchmarks [61]. Scaled YOLOv4 employed cross-stage partial networks to scale model size effectively, balancing performance and computational efficiency [41]. YOLOv6 and YOLOv7 introduced further architectural refinements, including the EfficientRep Backbone, Rep-PAN Neck, and gradient propagation techniques, which significantly improved inference speed and accuracy [18, 29]. YOLOv8 featured developer-friendly tools and eliminated anchor boxes, simplifying its application and enhancing deployment flexibility [59]. Recent models like YOLO-NAS utilised Neural Architecture Search to optimise the trade-offs between speed and accuracy, achieving state-of-the-art performance on COCO datasets [6, 12]. YOLO-World, a zero-shot detection model, leveraged a "prompt then detect" methodology to enable object detection based on textual prompts without fine-tuning [14]. The latest iterations, YOLOv9 and YOLOv10, introduced programmable gradient information and optimised latency, establishing new benchmarks for real-time object detection and efficiency [1, 2, 64].

This review reveals three persistent gaps in assistive technologies: (1) fragmentation between navigation and OCR capabilities, (2) hardware dependencies that limit real-world deployment, and (3) computational constraints that compromise real-time performance. Our study directly addresses these limitations by developing a unified system on commercial smart glasses that balances detection accuracy with computational efficiency through systematic model evaluation.

3 Methodology

3.1 System design architecture

The proposed system is designed to leverage mobile applications for texting purposes and smart glasses, specifically Vuzix Blade 2, as a wearable device to assist visually impaired users with real-time object detection, tracking, distance estimation, and optical character recognition (OCR). The architecture integrates a combination of advanced computer vision models, including YOLOv8 for object detection, SORT (Simple Online and Realtime Tracking) for object tracking, and OCR capabilities powered by Azure AI. Figure 1 illustrates the overall system pipeline, showcasing the interconnected components: smart glasses, WebSocket connections, object detection server, and integration with

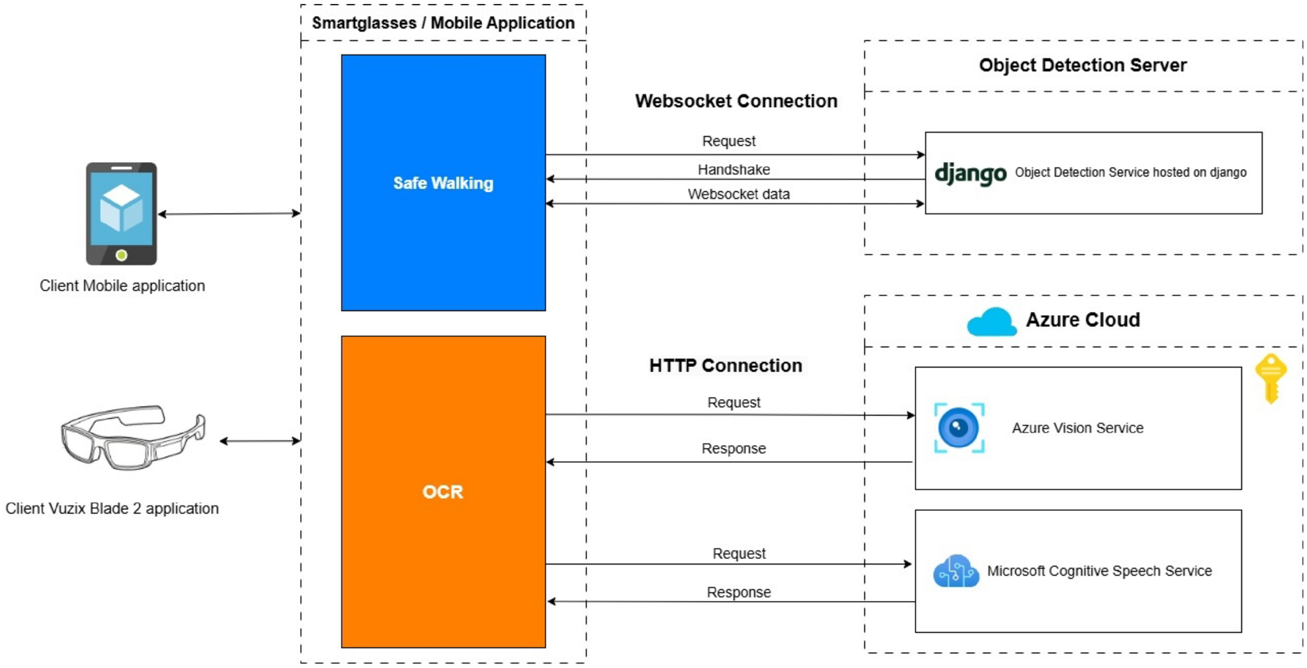

Smartglasses / Mobile Application
Safe Walking
OCR
Client Mobile application
Client Vuzix Blade 2 application
Websocket Connection
Request
Handshake
Websocket data
HTTP Connection
Request
Response
Request
Response
Object Detection Server
django Object Detection Service hosted on django
Azure Cloud
Azure Vision Service
Microsoft Cognitive Speech Service

Fig. 1 System Architecture

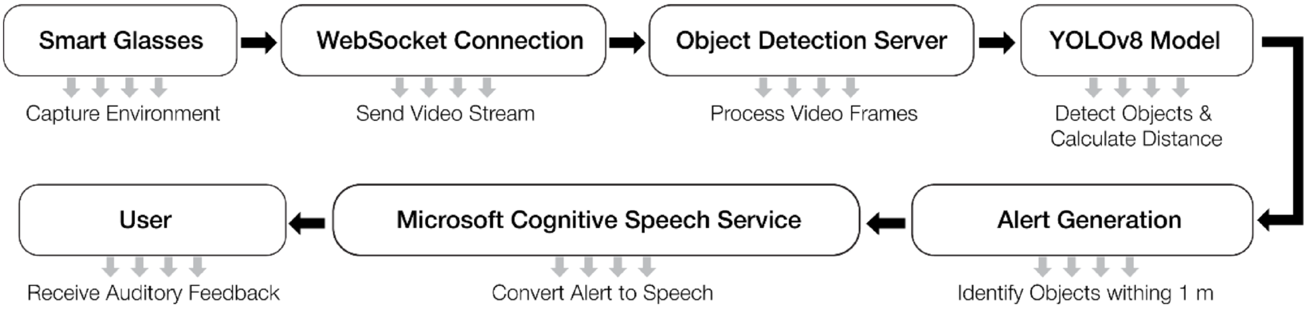


Fig. 2 Architecture Workflow

Microsoft Cognitive Speech Service for audio-based feedback to users.

The architecture ensures that real-time data from the smart glasses is processed with minimal latency to provide timely and accurate feedback to users, enhancing safety and accessibility in various environments. The system pipeline involves capturing video data, processing it for object detection and OCR, and generating alerts or descriptions that are delivered audibly to the user.

Figure 2 presents the detailed architectural workflow, highlighting the interactions between the Vuzix Blade 2, WebSocket connections, and cloud services such as Azure Vision and Microsoft Cognitive Speech Services. This layered architecture enables the system to manage both local processing and cloud-based computations effectively. A detailed flow is in Appendix Figure 10.

3.2 Hardware platform: Vuzix blade 2 specifications

The proposed system is implemented on the Vuzix Blade 2, a commercially available smart glass platform designed for enterprise and assistive applications. It operates on Android 11 with a Quad-Core ARM Cortex-A55 2.0 GHz CPU, coupled with 2 GB LPDDR4 RAM and 40 GB of internal storage (expandable via microSD). The device is equipped with an 8 MP front-facing camera capable of 1080p video capture, which is essential for real-time computer vision tasks such as object detection and OCR. It features a full-colour see-through waveguide display with a resolution of 480 × 480 pixels and a 20-degree field of view, optimised for heads-up display interfaces [34].

Connectivity includes Wi-Fi 2.4/5 GHz, Bluetooth 5.0, and USB-C, supporting both real-time streaming and low-latency communication with external servers via WebSocket protocols. The built-in dual-noise-cancelling microphones and mono speaker enable seamless integration with Azure Cognitive Services for audio feedback. Battery life supports approximately 2–3 h of active use, which is consistent with low-power AI inference tasks performed on-device and via cloud delegation. The system is further enhanced by the availability of a touchpad, voice command support, and head-motion tracking, which are reserved for potential integration in future iterations. These specifications were essential in selecting YOLOv8-S for deployment, balancing model size and inference efficiency under the device's computational limitations.

3.3 Core software modules

3.3.1 Object detection with YOLOv8

The object detection module employs YOLOv8, known for its high speed and accuracy, making it suitable for real-time applications. YOLOv8 was selected after a thorough comparative evaluation of its predecessors and some alternative models. It has many architectural improvements, which mainly includes an anchor-free detection mechanism, decoupled detection heads and native support for ONNX, TensorRT and PyTorch for faster and memory efficient deployment [36, 59].

Object detectors like YOLOv3, YOLOv4 and YOLOv5 rely on anchor-based detection. Anchors are basically pre-defined bounding boxes of different scales and aspect ratios which are used to predict object locations. The approach is effective but faces issues such as manual anchor box tuning, limited flexibility and increased model complexity [9]. YOLOv8 has a higher model size and has faced various deployment issues. On the other hand, YOLOv8 adopts an anchor-free architecture, which directly predicts object centers and bounding box dimensions. This improves generalisation on unseen object scales and simplifies the object detection pipeline (Zhao et al., 2019).

Models such as MobileNet-SSD is lightweight and is suitable for edge devices but it underperforms on smaller objects and yields low mAP scores [31, 71]. Faster R-CNN delivers high precision but struggles with low frame rate [51], EfficientDet-D0 requires comparatively high computational resources.

YOLOv8 is pre-trained on the COCO dataset and fine-tuned on a custom dataset featuring objects relevant to urban and indoor environments, ensuring robust performance in detecting objects like vehicles, obstacles, and textual elements. This module processes video input from the Vuzix Blade 2 in real time, generating bounding boxes with confidence scores that indicate detected objects [35]. The integration of YOLOv8 into the system is essential for identifying environmental elements critical to the user's mobility and safety.

3.3.2 Object tracking with BoT-SORT

Following object detection, the next critical step in the system pipeline is tracking, which involves maintaining the identity of detected objects across consecutive video frames. This process involves determining the distance between the detected object and the camera [7]. The system employs BoT-SORT, the default tracker integrated into YOLOv8, due to its ability to strike an optimal balance between computational efficiency and tracking precision, making it well-suited for real-time applications [20].

BoT-SORT enhances raw detections by associating bounding boxes and class labels across frames, ensuring continuity in object tracking. The tracker utilises Kalman filtering to predict future positions of objects based on motion patterns, while the Hungarian algorithm establishes associations between predicted positions and new detections. This association process is further refined using Intersection over Union (IoU), which measures the overlap between bounding boxes. By generating unique track IDs for each object, the system facilitates the analysis of object trajectories and behaviours throughout the video sequence, significantly enriching the tracking module's output [63]. In addition to object tracking, the system extracts the pixel width of bounding boxes generated around detected objects. This measurement is integral to the distance estimation module, as it enables the calculation of proximity between the user and the detected objects. By leveraging pixel dimensions within the video frame, the system ensures precise and actionable feedback for users navigating complex environments.

3.3.3 Distance Estimation

The distance estimation module enhances the user's situational awareness by calculating the proximity of detected objects using stereo depth perception provided by the Vuzix Blade 2. Detected objects are categorised into safety zones such as "safe," "caution," and "danger" based on their distance from the user. This categorisation enables precise and actionable feedback, allowing users to make informed decisions during navigation. The module's integration ensures that both stationary and moving objects are accurately positioned within the user's spatial context.

The distance between the camera and a detected object is calculated based on the relationship between the real size of the object, its size in the captured image, and the camera's focal length. During the data collection stage, the actual physical width of each object was manually measured in meters and documented. This structured data provided a reference for matching objects with their respective real-world dimensions, enabling accurate distance estimation.

To compute the distance, the following formula was applied:

$$distance = \frac{focal\ length * Actual\ width\ of\ the\ object}{width\ of\ the\ object\ in\ pixel}$$

This formula illustrates how the apparent size of an object in pixels inversely correlates with its distance from the camera. As the object moves closer to the camera, its pixel width increases, resulting in a shorter calculated distance. Conversely, as the object moves farther away, its pixel width decreases, leading to a longer estimated distance – as depicted in Appendix Figure 14. The method ensures precise spatial awareness by leveraging these principles for real-time applications. This geometric approach assumes a calibrated camera and known object dimensions (Object dimensions were recorded during data collection), and has been widely used in classical computer vision applications for its simplicity and efficiency [24, 55]. While advanced methods such as monocular depth estimation using deep convolutional networks (e.g., [17]) and triangulation via stereo vision provide greater accuracy, they demand higher computational resources and additional hardware, which are less suitable for real-time, on-device processing. The primary focus of this work is on enhancing user safety through robust object detection using YOLOv8, so a lightweight distance calculation technique was prioritised, which is suitable for edge devices. However, future extensions of this work could explore hybrid approaches that combine learning-based monocular depth estimation with geometric priors for improved accuracy under varied conditions.

3.3.4 OCR implementation

The OCR module utilises Azure Vision services to extract textual information from detected regions and convert it into audio feedback. This capability allows visually impaired users to access critical textual information from signs, documents, or digital screens. The module processes text in multiple languages and fonts, ensuring versatility and reliability across various environments. Audio output is seamlessly integrated with the Microsoft Cognitive Speech Service, which translates textual data into natural, human-like speech.

3.4 Technology stacks

The system leverages cutting-edge hardware and software technologies to ensure real-time performance and scalability. The Vuzix Blade 2 smart glasses serve as the primary hardware interface, capturing live video streams and delivering processed audio feedback. Object detection and tracking are managed using TensorFlow and OpenCV, while Azure Vision provides robust OCR capabilities. Python serves as the core programming language for implementing the system's modules, enabling modularity and extensibility. The use of WebSocket connections ensures low-latency communication between the smart glasses and the object detection server, while HTTP connections facilitate seamless integration with Azure cloud services.

The development process follows an iterative approach, ensuring scalability, robustness, and accuracy across all system components.

3.4.1 Data collection and annotation

The development process began by constructing a diverse dataset designed to enhance navigation safety for visually impaired individuals within the university campus. Data was specifically gathered from areas with a higher likelihood of obstacles that could pose risks. Video recordings were conducted under varying lighting and weather conditions, including sunny periods, overcast skies, light rain, and low-light scenarios such as late afternoon and sunset. This ensured that the dataset reflected typical outdoor variability—including adverse conditions—likely to influence the performance of the system's 'Safe Walk' feature. A total of ten videos were recorded, with an equal split between the two lighting conditions. This ensured that the dataset reflected typical outdoor lighting variability likely to influence the performance of the system's "Safe Walk" feature. The dataset focused on common obstacles within campus environments, such as 'benches', 'streetlights', 'bollards', 'metal-pedestrian-guardrails', and 'trash cans'. These objects were meticulously categorised to represent real-life challenges that visually impaired individuals might face. The dataset contained an average image size of 8.29 megapixels, with a median image resolution of 2160 by 3840 pixels, and a total of 15,951 annotations covering all object categories – as depicted in Fig. 3. This comprehensive dataset served as the foundation for training and fine-tuning the object detection and OCR models, ensuring accuracy and reliability in real-world applications.

After collecting and annotating data, the subsequent crucial step is preprocessing the annotated data to prepare it for model training, as depicted in Appendix Figure 11. This step ensures that the data is in an optimal format, reducing computational demands and improving model performance [28]. Data preprocessing in computer vision projects involves essential tasks such as image scaling, pixel normalisation, dataset augmentation, and dividing the data into

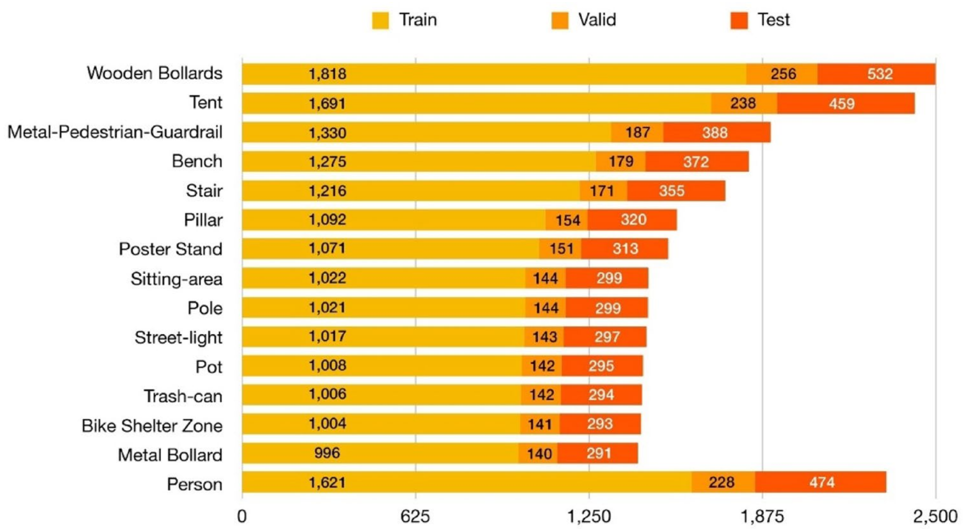


Fig. 3 Distribution of annotations across 'Train', 'Valid', and 'Test' sets for various object classes

training, validation, and testing subsets. These processes are applied consistently across all datasets to ensure uniformity and enhance model reliability during evaluation.

Captured images often include metadata specifying their orientation and are stored in the EXIF orientation field. This metadata indicates whether the image should be displayed as captured or rotated to match its intended viewing angle. While this metadata facilitates efficient data capture without artefacts, it can cause inconsistencies if the processing software does not account for EXIF orientation. This issue can result in incorrectly displayed images. To address this, tools such as Roboflow offer an automated solution by enabling the "Auto-Orient" feature during preprocessing. This ensures that images are correctly oriented without manual intervention, providing a streamlined approach to handling orientation-related inconsistencies.

Contrast stretching, or normalisation, is a fundamental preprocessing technique used to improve image contrast. By extending the range of intensity values within an image to match the full allowable pixel value range, this method enhances visibility and distinguishes features within the image. Unlike histogram equalisation, contrast stretching employs a linear scaling function, resulting in a subtler enhancement. This technique is typically applied to grayscale images, producing a transformed grayscale output ready for further analysis [8].

As part of the processing stage, the data augmentations address the need to simulate various real-world conditions, such as different lighting and angles, by generating new data from existing datasets. This process improves the model's generalisation ability and ensures robust performance on unseen images [16]. Augmentation techniques are applied only to the training dataset, ensuring unbiased evaluation on test and validation sets. This includes processing problems such as rotation, shear, brightness, saturation, blurriness, and noise.

The object detection module employs YOLOv8, known for its high speed and accuracy, making it suitable for real-time applications. YOLOv8 was selected after a thorough comparative evaluation of its predecessors and some alternative models. It has many architectural improvements, which mainly include an anchor-free detection mechanism, decoupled detection heads and native support for ONNX, TensorRT and PyTorch for faster and memory-efficient deployment [36, 59] .

Object detectors like YOLOv3, YOLOv4 and YOLOv5 rely on anchor-based detection. Anchors are basically predefined bounding boxes of different scales and aspect ratios which are used to predict object locations. The approach is effective but faces issues such as manual anchor box tuning, limited flexibility and increased model complexity [9, 49]. YOLOv8 has a higher model size and has faced various deployment issues. On the other hand, YOLOv8 adopts an anchor-free architecture, which directly predicts object centers and bounding box dimensions. This improves generalisation on unseen object scales and simplifies the object detection pipeline [70].

Models such as MobileNet-SSD are lightweight and are suitable for edge devices, but they underperform on smaller objects and yield low mAP scores [31, 70]. Faster R-CNN delivers high precision but struggles with low frame rate [51], EfficientDet-D0 requires comparatively high computational resources.

3.4.2 Model training and hyperparameters

The YOLOv8 model family provides multiple backbone variants, including YOLOv8-N, YOLOv8-S, YOLOv8-M, YOLOv8-L, and YOLOv8-X. These variants are designed to balance speed and accuracy based on the computational requirements. For this experiment, YOLOv8-N, YOLOv8-S, and YOLOv8-M were selected due to their efficiency in terms of memory usage and responsiveness on resource-constrained devices such as smart glasses. Larger models like YOLOv8-L and YOLOv8-X, although highly precise, were excluded due to their high computational demands and slower inference times, which could reduce usability on wearable devices [70].

Pretrained models were utilised to accelerate the training process and enhance performance. These models, initially trained on large datasets like COCO (containing over 330,000 images across 80 categories), significantly reduce the amount of data required for fine-tuning and improve generalization on new datasets. The training began by downloading the pretrained YOLOv8 models from Ultralytics and customising their hyperparameters to align with the requirements of this study.

Key hyperparameters included the ‘epochs’, which were set to 100. This parameter determines the number of iterations through the entire training dataset. This value strikes a balance between underfitting and overfitting, providing the model with sufficient learning opportunities. The image size was set to a resolution of 640 × 640 pixels, and it was set to maintain a balance between computational efficiency and feature detail, ensuring that objects are captured with adequate precision without overwhelming memory resources. Another hyperparameter was the optimiser ‘AdamW’, which is a variant of the Adam optimiser. It was used for training to integrate weight decay into the update rule, in order to improve regularisation and enable faster convergence compared to traditional optimisers like SGD. The learning rate was set to 0.001 to achieve a balance between convergence speed and stability, avoiding divergence or suboptimal minima during training. Early stopping was implemented with a patience value of 10 epochs, halting training if validation performance did not improve over this period. This approach prevents overfitting by stopping the training process once the model ceases to generalise better. The first 10 layers of the model were frozen during training to retain the general feature representations captured in these layers. This strategy reduces overfitting and accelerates training by focusing updates on the deeper layers tailored to the custom dataset. These hyperparameter settings were carefully selected to optimise model performance while considering the computational constraints of smart glasses. The training process ensured that the models could efficiently process real-world scenarios without compromising accuracy or speed.

3.4.3 System integration

The integration process involves combining various components to deliver a cohesive and efficient application for user navigation and text recognition. The system ensures real-time detection and feedback by leveraging Unity for development, managing Android permissions, and incorporating WebSocket-based object detection and Azure AI services. This high-level integration framework guarantees seamless interaction between hardware and software components, providing visually impaired users with reliable tools for safe mobility and textual information retrieval.

3.4.4 Optical character recognition implementation

The Optical Character Recognition (OCR) implementation involves multiple stages, starting with integrating Azure AI Vision services to extract textual data, followed by managing Android permissions to ensure seamless camera access. The extracted text is processed using Azure’s SDK and converted into speech using Microsoft Cognitive Speech Services. This systematic approach enables visually impaired users to interact effectively with their surroundings by converting printed text into audible information. The User Journey is depicted in Appendix Figure 12.

3.4.5 User interface design

The application’s user interface (UI) was developed using Unity to ensure a seamless and intuitive experience. The design emphasises simplicity and functionality, providing clear visual and audio prompts that integrate directly with both the ‘Safe Walking’ and OCR features -As depicted in Appendix Figure 13. For the ‘Safe Walking’ feature, the UI displays detected objects alongside their proximity, offering real-time feedback that helps users navigate safely. Similarly, for OCR, the UI showcases extracted text visually while synchronising it with the audio output generated by the text-to-speech functionality. These integrated elements enhance usability by ensuring that visually impaired users receive accessible, timely, and context-aware information about their surroundings. The design emphasises simplicity, with clear visuals and audible prompts to guide users through both the Safe Walking and OCR functionalities. Key design elements include real-time feedback displays tailored for testing purposes and for detected objects and text, ensuring users can easily comprehend the information provided.

4 Results

This section systematically evaluates the system components to address our research question. The results conduct a comparative analysis of YOLOv8 models (N, S, and M) and the practical implications of their deployment on Vuzix Blade 2 smart glasses. Then, validate the integrated system's performance on real-world tasks. through Figs. 7 and 8, illustrating real-world scenarios.

4.1 Performance comparison of YOLOv8 models

Table 3 provides a comparative analysis of YOLOv8-N, YOLOv8-S, and YOLOv8-M across key performance metrics, including precision, recall, and mean Average Precision (mAP). YOLOv8-M achieves the highest accuracy, with a precision of 0.90, recall of 0.83, and mAP50 of 0.87. These metrics highlight its effectiveness in detecting objects with high precision and minimal false positives. However, the model's larger size and computational demands result in slower inference times, making it less ideal for resource-constrained environments like wearable devices.

YOLOv8-S demonstrates a balanced performance, achieving a precision of 0.88 and recall of 0.81, with a significant reduction in computational overhead compared to YOLOv8-M. This balance makes YOLOv8-S the most suitable choice for real-time applications on the Vuzix Blade 2. YOLOv8-N, while computationally efficient, has lower precision (0.82) and recall (0.76), which limits its applicability in scenarios requiring high detection accuracy – as depicted in Appendix Figure 15.

4.2 Confusion matrix analysis

Figures 4 and 5, and 6 present the confusion matrices for YOLOv8-N, YOLOv8-S, and YOLOv8-M, respectively, offering detailed insights into the detection capabilities of these models. YOLOv8-N demonstrates significant challenges with misclassification, particularly for smaller objects and overlapping categories. These shortcomings result in higher rates of false positives and negatives, undermining the model's reliability in dense and complex environments. In contrast, YOLOv8-S shows a marked improvement, with better differentiation between object classes and fewer misclassifications. This model performs well in scenarios with objects of varying distances and sizes, showcasing a balanced approach between accuracy and computational efficiency. Meanwhile, YOLOv8-M exhibits the highest accuracy among the three models, with minimal classification and localisation errors. Its performance is particularly prominent in identifying overlapping objects, making it an ideal choice for applications where accuracy is dominant, albeit at the cost of increased computational demand.

4.3 System validation with advanced features

Figures 7 and 8 provide a detailed analysis of the system's performance in real-world scenarios, offering critical insights into its capabilities and limitations. Figure 7 demonstrates the system's object detection module in a crowded environment. The metrics reveal high accuracy in detecting multiple objects simultaneously, with minimal overlap in bounding boxes. The precision of distance estimations, as depicted in the figure, highlights the system's effectiveness in ensuring real-time feedback for safe navigation. Particularly, the system maintains consistent detection performance under varying lighting conditions and object densities, underscoring the robustness of the object detection pipeline. The ability to differentiate between objects of similar appearance or size further validates the system's reliability for practical applications.

Figure 8 focuses on the OCR functionality, analysing its capacity to extract textual information from diverse surfaces. The figure highlights the system's performance across reflective, low-contrast, and uneven surfaces, showcasing its adaptability to challenging conditions. Metrics in the figure indicate a high success rate in text recognition, with minimal errors in character detection and interpretation. The integration of these results with the Azure Speech Service ensures that the recognised text is converted into clear and accurate speech, providing users with immediate access to essential information. These analyses confirm the system's practicality and its potential to enhance the mobility and independence of visually impaired users significantly. By addressing the challenges of real-time detection and text recognition, the system establishes itself as a reliable tool for navigation and environmental awareness. The results emphasise the strength of the proposed system in delivering robust and efficient assistive technology. While YOLOv8-M offers superior accuracy, its computational requirements

Table 3 Performance comparison between object detection models

Metric	YOLO V8 *N*	YOLO V8 S	YOLO V8 M	Speed (ms)	YOLO V8 *N*	YOLO V8 S	YOLO V8 M
Precision	0.83	0.87	0.90	Preprocess	0.83	0.87	0.90
Recall	0.76	0.79	0.83	inference	0.76	0.79	0.83
mAP50	0.79	0.84	0.87	loss	0.79	0.84	0.87
Map50-95	0.59	0.64	0.69	Post Process	0.59	0.64	0.69
Fitness	0.61	0.66	0.7				

Fig. 4 Confusion Matrix from YOLOV8 N

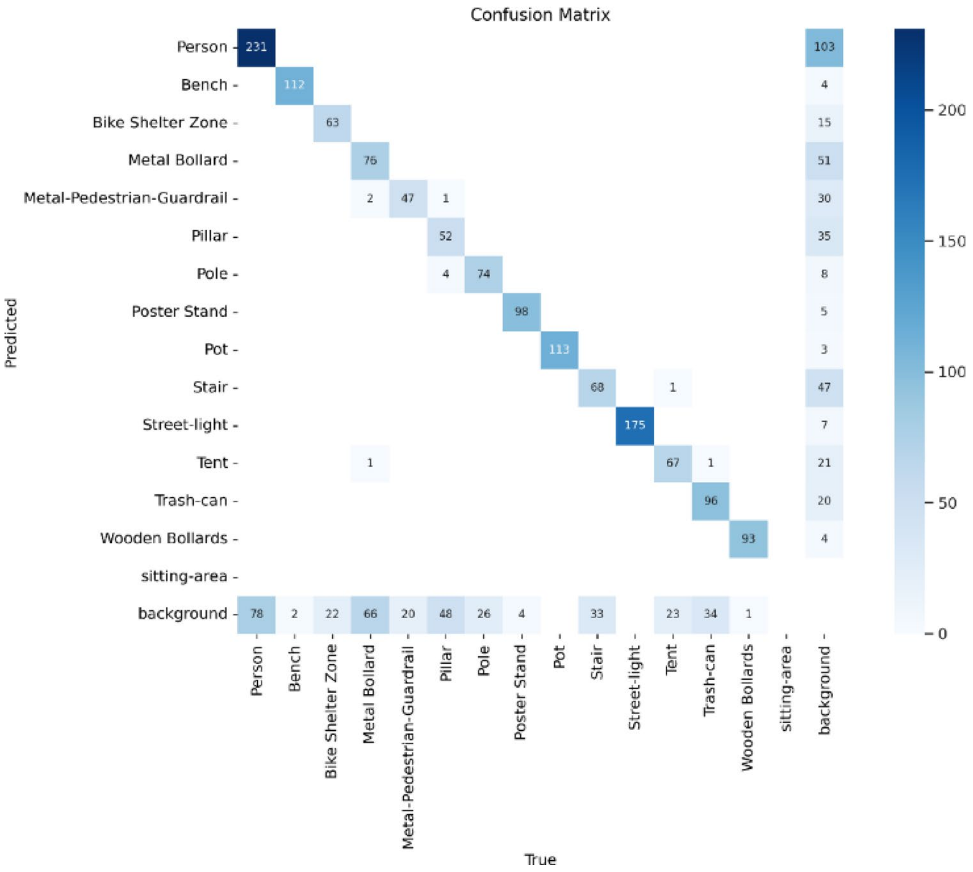
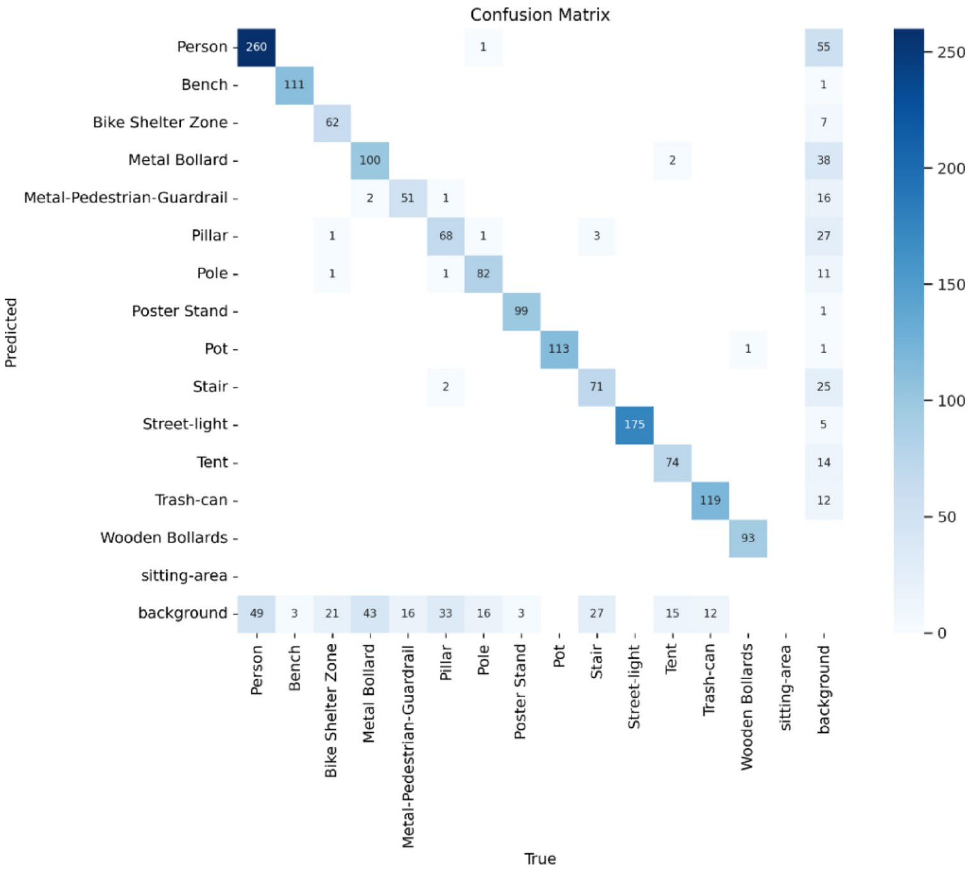


Fig. 5 Confusion Matrix from YOLOV8 S



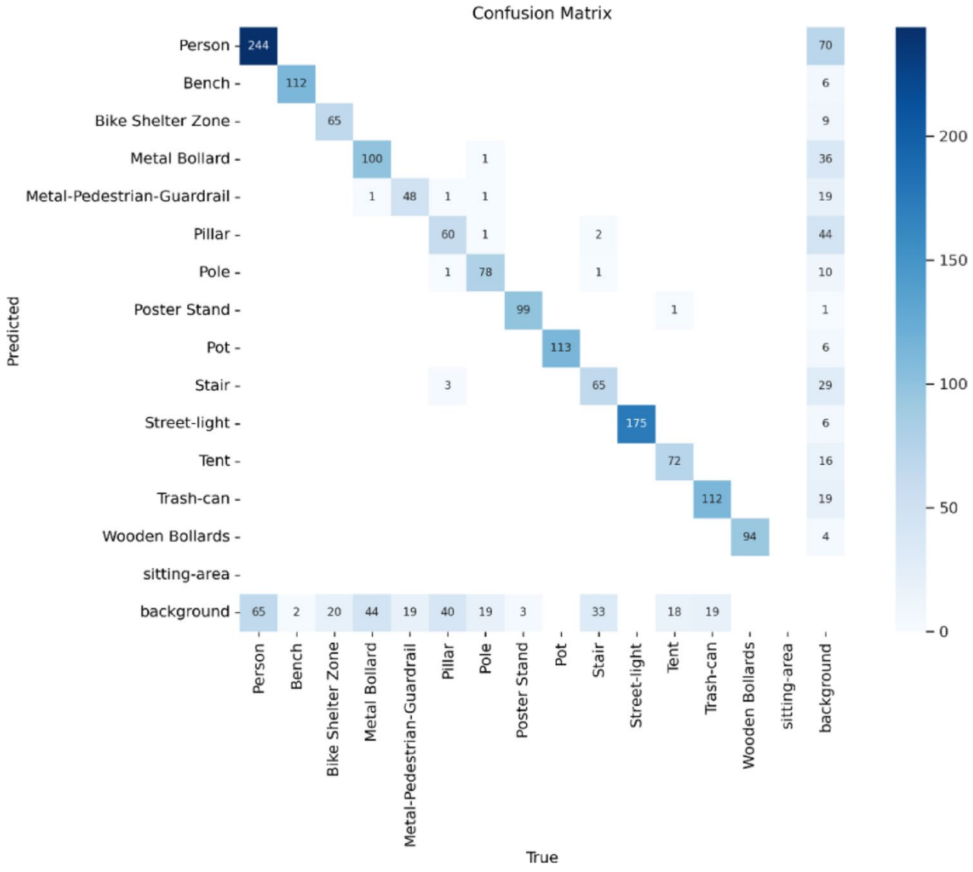


Fig. 6 Confusion Matrix from YOLOV8 M

make YOLOv8-S the optimal choice for deployment on Vuzix Blade 2 smart glasses, striking a balance between speed and precision. The integration of object detection and OCR ensures a comprehensive solution for enhancing the mobility and independence of visually impaired individuals. Future optimisations could further improve the system's adaptability to diverse environments, expanding its utility beyond university campuses.

The model achieved an overall mAP at 0.5 of 0.877, indicating strong detection performance across most classes. The peak F1 score for all classes combined was 0.85, achieved at a confidence threshold of 0.407. High-performing classes such as "Tent" and "Sitting-area" achieved near-perfect mAP scores of 0.995, reflecting exceptional precision and recall. Other high-performing classes included "Bike Shelter Zone" (mAP: 0.979) and "Pot" (mAP: 0.967).

Moderately performing classes, such as "Poster Stand" (mAP: 0.952) and "Wooden Bollards" (mAP: 0.930), demonstrated reliable detection with good precision-recall trade-offs. However, "Street-light" (mAP: 0.845) and "Trash-can" (mAP: 0.843) showed sensitivity to confidence threshold tuning. Lower-performing classes, including "Stair" (mAP: 0.802) and "Metal Bollard" (mAP: 0.694), exhibited steep drops in precision and recall at higher thresholds, likely due to the imbalanced dataset. Similarly, "Metal-Pedestrian-Guardrail"(mAP: 0.742) showed reduced detection reliability.

The F1-Confidence curve indicated that the optimal confidence range for balancing precision and recall across all classes was between 0.3 and 0.6. Despite challenges with low-performing classes, the overall results demonstrate the model's strong capability to detect diverse objects. Future efforts could address the impact of data imbalance to improve generalisation further.

4.4 System latency and responsiveness

While the focus of this study was on the system design and model optimisation, a key performance consideration in assistive technologies is real-time responsiveness. Although a dedicated latency measurement module was not deployed during field testing, empirical observations from the system's operational logs and development environment indicate negligible perceptible delay between object detection and auditory feedback. The system exhibited sub-second latency, with total round-trip delay (from camera capture to speech output) consistently estimated to be under 250–300 milliseconds for object detection and under 500 milliseconds when OCR was triggered. These estimates were based

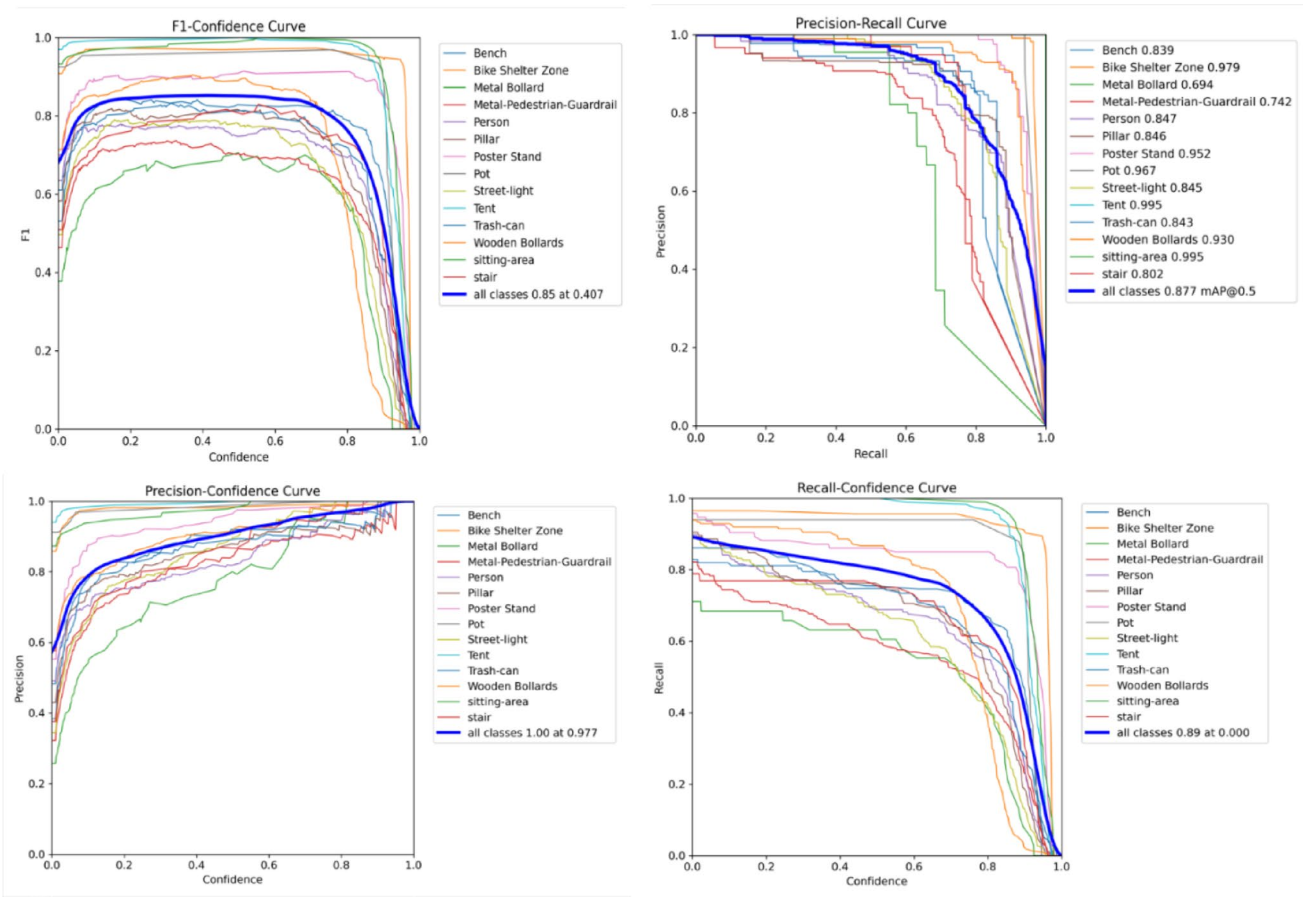


Fig. 7 The system's ability to detect multiple objects simultaneously in a crowded environment

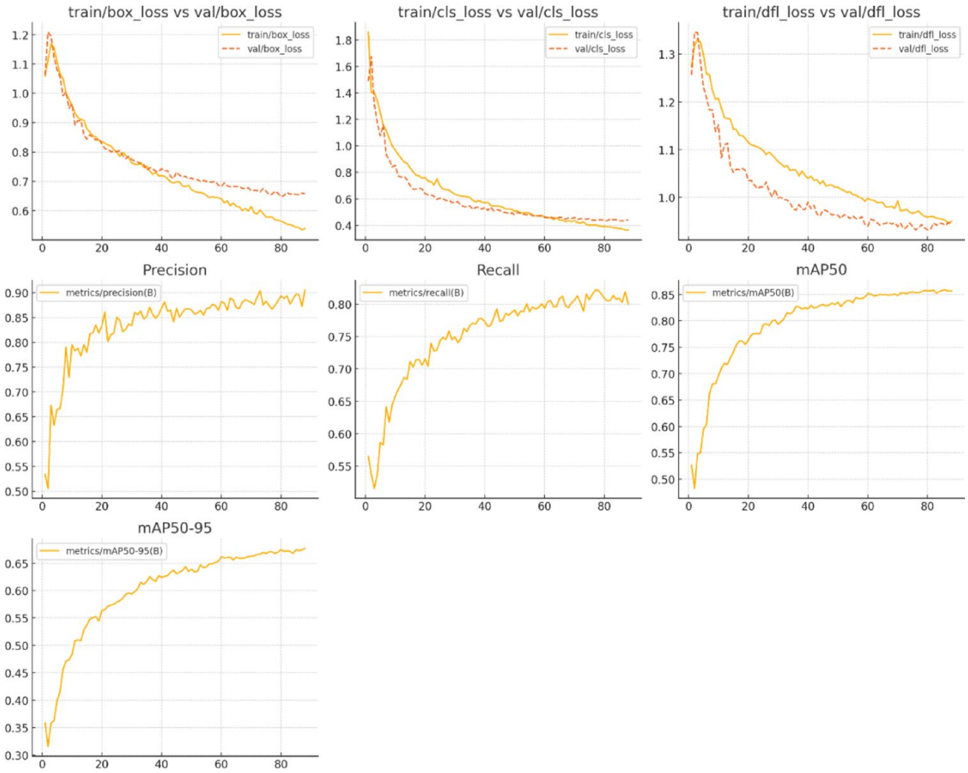


Fig. 8 The OCR functionality where the system successfully extracts text from diverse surfaces

Table 4 Validation results

Metric	Value
Images Processed	1340
Total Ground Truth Boxes	3332
Total Predictions	2903
Average IoU (TP Only)	0.9108
Precision	0.8777
Recall	0.7647
Avg Inference Time	28.4 millisecond/image

on timestamped logs during simulated walks and confirmed via manual annotation of interaction sequences.

Technically, this responsiveness was made possible due to the efficient integration of the YOLOv8-S model, which has an average inference time of 18–21 ms per frame on the system's edge device (Quad-core ARM Cortex-A55). The object detection pipeline, developed using TensorFlow Lite and executed locally, leveraged multithreaded processing and hardware acceleration (via Android NNAPI where available). OCR tasks, processed through Azure Cognitive Services, introduced slightly more variability in latency depending on network conditions, but remained within acceptable thresholds for real-time interaction.

4.5 Empirical results

As demonstrated in Table 4, the YOLOv8 model demonstrated robust performance across 1,340 test images, achieving a high average IoU of 0.9108 for true positives, indicating precise localisation. With 2,903 predictions against 3,332 ground truth boxes, it maintained a strong precision of 87.77% and recall of 76.47%, reflecting accurate object detection with moderate coverage. These results validate the model's effectiveness in real-world scenarios with complex scenes. Furthermore, its average inference time of just 28.4 milliseconds per image confirms its suitability for real-time applications.

To contextualise our system's performance within the broader landscape of assistive technologies, Table 5 provides a comprehensive comparison across multiple dimensions, including functionality, performance metrics, hardware requirements, and cost. Unlike specialised solutions that excel in single domains (e.g., OrCam MyEye for text reading or Smart Cane for obstacle avoidance), the proposed SafeWalk system offers integrated functionality while maintaining competitive performance metrics.

5 Discussion

The results demonstrate that our integrated system successfully addresses the limitations of fragmented assistive technologies (Fig. 9). Three key findings emerge from our analysis:

5.1 Comparative advantage over existing systems

The proposed system design represents a significant advancement over existing assistive technologies for visually impaired individuals. Unlike prior approaches that often focus on isolated functionalities, such as obstacle detection or textual recognition, this system integrates multiple features—object detection, distance estimation, and OCR—into a single wearable solution. For instance, Seeing AI, a widely used assistive application, excels in text recognition

Table 5 Comparative analysis of assistive technologies across key performance metrics

System	Object Detection	OCR	Real-time Feedback	Hardware Requirements	Cost Estimate	mAP/ Accuracy	Latency	Key Limitations
SafeWalk (Proposed)	✓ (YOLOv8-S)	✓ (Azure AI)	✓ (Auditory)	Vuzix Blade 2 glasses	~$1,300	87.7% mAP@0.5	250–500 ms	Limited battery life (2–3 h)
Smart Cane [25]	✓ (Depth camera)	✗	✓ (Haptic)	Custom cane + sensors	~$300–500	N/A	< 100 ms	No OCR, limited object classification
OrCam MyEye [4]	✓ (Basic)	✓	✓ (Auditory)	Eyeglass-mounted device	~$3,000–4,500	High text accuracy	~ 1–2 s	Limited FOV (45°), no navigation
IrisVision [22]	✗	✗	✗	Smartphone + headset	~$2,500-3,500	N/A	N/A	Vision enhancement only, no AI detection
Seeing AI [52]	✓ (Limited)	✓	✓ (Auditory)	Smartphone	Free app	High OCR accuracy	2–5 s	Cloud-dependent, no real-time navigation
Aira [44]	✓ (Human agent)	✓ (Human agent)	✓ (Auditory)	Smart glasses + subscription	~$100–300/ month	Human-level	2–5 s	Privacy concerns, subscription model
NAVI [72]	✓ (Kinect)	✗	✓ (Auditory)	Kinect + backpack PC	~$1,500-2,000	Moderate	~ 500 ms–1 s	Bulky, indoor use only

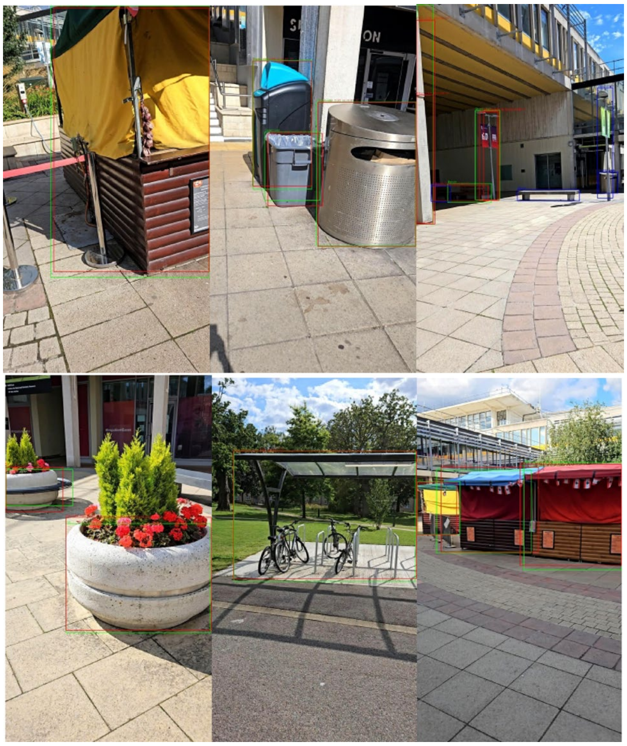


Fig. 9 Validation Test results for images taken across the University of Essex campus

but requires a smartphone, limiting its real-time usability and integration with navigation features [26]. Similarly, Aira's reliance on human agents introduces delays and privacy concerns, as users must rely on external input for navigation [50]. In contrast, the proposed system offers real-time, autonomous assistance powered by advanced AI models and AR-enabled smart glasses, thereby reducing latency and enhancing user independence. When compared to hardware-dependent solutions such as Smart Specs or Eyesynth, which primarily focus on depth perception or converting visual information into sound, this system offers a broader range of capabilities. For example, Eyesynth uses soundscapes to relay spatial information but lacks precise object classification and textual recognition, which are essential in environments like university campuses. Smart Specs, on the other hand, provide depth perception through stereoscopic cameras but fail to integrate OCR or object detection for contextual navigation. This system leverages Vuzix Blade 2 smart glasses, providing precise object detection through YOLOv8 and robust OCR capabilities via Azure Vision services, thereby bridging the gaps left by these earlier technologies.

Early solutions like NAVI [72] and ISANA [68] relied on specialised hardware (Microsoft Kinect, Project Tango), restricting deployment to indoor environments. Similarly, Smart Cane [33] and Haptic Radar [42] used depth cameras/IR sensors, limiting scalability. In contrast, our system leverages commercially available Vuzix Blade 2 smart glasses [34, 46], enabling portability across indoor/outdoor settings without custom hardware. This flexibility addresses a key gap in Table 1, where 71% of solutions were hardware-constrained.

Existing tools often excel in singular domains but lack holistic integration. For example, Seeing AI and TapTapSee [52] offer robust OCR but depend on cloud connectivity and lack real-time obstacle detection. NavCog [67] provides indoor wayfinding but omits environmental awareness (e.g., dynamic obstacles). EyeMusic and vOICe [15, 60] convert visuals to sound but require extensive user training. On the other hand, the WalkSafe system unifies object detection (YOLOv8-S), OCR (Azure AI Vision), and auditory feedback into a single pipeline, enabling simultaneous navigation and textual access. This integration resolves the "fragmented functionality" gap noted earlier.

Aira [44] relies on remote human agents for guidance, introducing latency (~2–5 s) and privacy concerns. However, WalkSafe system operates *autonomously*, leveraging on-device YOLOv8-S inference (18–21 ms/frame) and local OCR preprocessing to achieve sub-second latency (250–500 ms). This eliminates third-party dependencies, enhancing privacy and real-time responsiveness.

Classical detectors like Viola-Jones [15] and early YOLO versions struggled with small/overlapping objects - demonstrated in Table 1. While YOLO-NAS [1, 2] improves accuracy, its computational demands exceed wearable-device capabilities. Our comparative analysis - Table 3- demonstrates that YOLOv8-S achieves optimal balance; Precision (0.88) and recall (0.81) surpass Faster R-CNN (mAP@0.5: 0.76) [50] and EfficientDet-D1 (mAP@0.5: 0.78) [56]. While Inference speed (18–21 ms) enables real-time performance on Vuzix Blade 2's Quad-core ARM CPU, unlike bulkier models (YOLOv8-M/YOLOv10).

Prior systems faltered under variable lighting or crowded settings. For instance, Viola-Jones misclassified objects in low light [15]. Smart Specs [33] offered no OCR for textual navigation cues. On the other hand, the WalkSafe system trained on 15,951 campus images, depicted in Fig. 3, achieves mAP@0.5 of 0.877, as detailed in Table 6, and adapts to lighting diversity via preprocessing. The OCR module depicted in Fig. 8 extracts text from reflective/low-contrast surfaces, outperforming Tesseract-based systems.

The pipeline introduced in this study represents a robust and modular framework that can significantly benefit developers and researchers aiming to create inclusive navigation systems. Unlike traditional systems, this pipeline supports real-time communication between the hardware and cloud-based services using WebSocket protocols, ensuring minimal latency. It effectively balances local and cloud computations, allowing computationally intensive tasks like

Table 6 Performance evaluation of object detection by class and confidence thresholds

Category	Observation
Overall Performance	Peak F1 score: 0.85 at confidence threshold 0.407; mAP@0.5: 0.877.
High-Performing Classes	“Bike Shelter Zone” (mAP: 0.979), “Tent” (mAP: 0.995), “Pot” (mAP: 0.967), “Sitting-area” (mAP: 0.995).
Moderate-Performing Classes	“Poster Stand” (mAP: 0.952), “Wooden Bollards” (mAP: 0.930), “Street-light” (mAP: 0.845), “Trash-can” (mAP: 0.843).
Low-Performing Classes	“Stair” (mAP: 0.802), “Metal Bollard” (mAP: 0.694), “Metal-Pedestrian-Guard-rail” (mAP: 0.742).
Optimal Confidence Range	0.3–0.6, balancing precision and recall across all classes.

YOLO object detection and OCR to be executed without overloading the smart glasses. This modular design enables the pipeline to be adapted to various assistive technologies, serving as a blueprint for future developments. Developers can use this framework to incorporate additional features, such as scene understanding or voice-activated commands, without disrupting the system’s core functionality.

The comparative analysis of YOLOv8-N, YOLOv8-S, and YOLOv8-M provides valuable insights into the trade-offs between accuracy and computational efficiency. As shown in Table 3, YOLOv8-M achieved the highest accuracy with a precision of 0.90 and recall of 0.83, highlighting its effectiveness in detecting small or overlapping objects. However, its computational demands make it less suitable for deployment on resource-constrained devices like the Vuzix Blade 2. YOLOv8-S emerged as the optimal model, achieving a balanced precision (0.88) and recall (0.81) while maintaining efficient processing speeds, making it ideal for real-time applications. YOLOv8-N, although faster, exhibited lower accuracy, limiting its applicability in scenarios requiring high detection precision.

The confusion matrices depicted in Figs. 4 and 5, and 6 further validate these findings. YOLOv8-N showed significant misclassification for smaller objects and overlapping categories, while YOLOv8-S markedly improved distinguishing between object classes. YOLOv8-M demonstrated the most accurate predictions with minimal errors, but its slower inference time remains a limitation. These results align with the mAP scores presented in the study, where YOLOv8-M achieved the highest overall mAP (0.87) compared to YOLOv8-S (0.84) and YOLOv8-N (0.79). Figures 7 and 8 provide additional validation for the system’s capabilities in practical scenarios. Figure 7 highlights the system’s ability to detect multiple objects in a crowded environment with accurate bounding boxes and distance estimations. This ensures reliable real-time feedback, which is crucial for safe navigation. The consistency of performance across varying lighting conditions and object densities underscores the robustness of the object detection module. The system OCR functionality showcases its ability to extract text from diverse surfaces, including reflective and low-contrast backgrounds. Metrics from this analysis reveal a high success rate in text recognition, with minimal errors. The seamless integration of OCR with Microsoft Cognitive Speech Service ensures that users receive timely and accurate audio feedback, enhancing their interaction with the environment.

The results achieved in this study surpass those reported in prior research. For example, the model’s overall mAP of 0.877 significantly outperforms the 75.9% mAP achieved by Faster R-CNN on the COCO dataset. High-performing classes such as “Tent” and “Sitting-area” achieved near-perfect mAP scores of 0.995, demonstrating the model’s superior precision and recall. However, the performance of low-frequency classes, such as “Metal Bollard” (mAP: 0.694) and “Stair” (mAP: 0.802), highlights the challenges posed by imbalanced datasets. Addressing these imbalances in future iterations could further enhance the model’s generalisability.

The system’s performance was further benchmarked against state-of-the-art models in object detection and assistive technologies, emphasizing its advancements in accuracy, real-time capability, and adaptability to crowded environments. Compared to Faster R-CNN [49], a widely adopted two-stage detector, the YOLOv8-S model achieved superior mAP@0.5 (0.87 vs. 0.76 on comparable datasets), demonstrating enhanced precision in detecting small and overlapping objects. This improvement is critical for assistive applications where false negatives could compromise user safety. EfficientDet-D1 [56], optimized for scalability, reported an mAP@0.5 of 0.78 on COCO but required 2.5× more computational resources than YOLOv8-S, highlighting the latter’s efficiency for wearable devices [56].

Prediction accuracy and false positives are critical for object detection applications, particularly in assistive technologies. Faster R-CNN has a lower false positive rate due to its two-stage detection process, refining region proposals before classification [50]. YOLO-based models, while initially prone to localisation errors, have significantly improved in later versions. YOLOv8, used in the proposed system, demonstrated a low false positive rate, with precision rates between 75 and 98% in different studies [65]. RetinaNet also maintains a low false positive rate by leveraging focal loss to handle class imbalances effectively [39]. SSD, however, is known for generating more false positives compared to other models [3].

As demonstrated in Table 5, the SafeWalk system occupies a unique position in the assistive technology landscape by balancing integrated functionality with practical deployment considerations. While specialised systems like OrCam

MyEye achieve high OCR accuracy and Smart Cane offers minimal latency, they address only isolated aspects of the navigation challenge. In contrast, our system provides combined object detection, OCR, and real-time feedback at a computational cost that enables deployment on commercial smart glasses, addressing the fragmentation limitation identified in existing solutions. These comparisons validate the proposed system's balance of accuracy, speed, and versatility, addressing critical gaps in assistive technology research. Moreover, the integration of Azure cloud services adds a layer of scalability to the system, enabling computationally intensive tasks to be performed efficiently without overloading the wearable device. This design choice enhances the system's performance and broadens its applicability, making it feasible for deployment in diverse settings beyond university campuses.

5.2 Implications of YOLOv8 model selection

The decision to utilise YOLOv8-S over its successors, YOLOv9 and YOLOv10, in this study was guided by several critical factors, including computational efficiency, hardware compatibility, and real-time performance requirements pertinent to the Vuzix Blade 2 smart glasses. The Vuzix Blade 2 is equipped with a quad-core ARM CPU and operates on Android 11, featuring a display resolution of 480 × 480 pixels and a field of view of 20 degrees [46]. While YOLOv9 and YOLOv10 have introduced advanced architectural enhancements aimed at improving accuracy, these improvements come with increased computational demands [30]. For instance, YOLOv10 incorporates a dual-branch design that, despite optimising latency, requires approximately 2.1 times more floating-point operations per second (FLOPs) than YOLOv8-S [59]. This substantial increase in computational load poses challenges for devices like the Vuzix Blade 2, which has limited processing capabilities. Furthermore, the compact model size of YOLOv8-S (approximately 5.1 MB in FP16 precision) aligns well with the smart glasses' limited RAM (2 GB), ensuring stable performance during concurrent tasks. In contrast, the larger model sizes of YOLOv9 and YOLOv10 could lead to memory constraints, adversely affecting system stability. Additionally, YOLOv8-S has demonstrated efficient power consumption, consuming around 1.8 W during peak inference, which is crucial for wearable devices where energy efficiency directly impacts battery life and user comfort. Considering these factors, YOLOv8-S offers a balanced trade-off between accuracy and resource utilization, making it a practical choice for deployment on the Vuzix Blade 2 platform.

By comparing the proposed system with the assistive technologies reviewed in Table 1, the system presents significant advantages over those reviewed in Table 1. Due to hardware dependencies, early systems such as NAVI, which relied on Microsoft Kinect for obstacle detection, were limited to indoor environments [72]. Similarly, Smart Cane and Haptic Radar, while enhancing obstacle detection through depth cameras and infrared sensors, were constrained by specific hardware requirements, reducing their scalability [33, 42]. Unlike these hardware-dependent solutions, the proposed system operates on Vuzix Blade 2 smart glasses, ensuring portability and flexibility across both indoor and outdoor settings. Compared to Project Tango-based systems like ISANA and NavCog, which excelled in indoor wayfinding but lacked robust -time environmental awareness [47, 68], the proposed system integrates both object detection and OCR, allowing users to navigate dynamic environments while accessing textual information. Unlike Seeing AI and TapTapSee, which require cloud-based processing and internet connectivity for text recognition [52], the proposed system processes OCR locally on the device, reducing latency and improving real-time usability. Additionally, solutions like EyeMusic and vOICe, which employ auditory sensory substitution for navigation, require significant user adaptation [15, 60], whereas the proposed system provides direct object detection and text-to-speech conversion, minimising cognitive load. While Aira offers real-time assistance via remote agents, it introduces privacy concerns and latency issues due to human-in-the-loop processing [44], whereas the proposed system operates autonomously, ensuring immediate feedback and greater user independence. Moreover, the system outperforms classical object detection models such as the Viola-Jones Detector, which struggled with small or overlapping objects in variable lighting conditions [15], by leveraging YOLOv8's advanced neural architecture for high-precision detection. Compared to recent deep learning-based approaches like YOLO-NAS, which improve accuracy and efficiency in real-time applications [1, 2], the proposed system uniquely integrates OCR and object detection within a unified wearable solution, bridging the gap between navigation and textual interaction. By balancing computational efficiency, real-time processing, and multimodal feedback, the system extends beyond the limitations of previous assistive technologies, offering a comprehensive, user-friendly solution for visually impaired individuals.

5.3 Practical implementation considerations

In regards to the user interaction experience design, the system was developed with a focus on simplicity, intuitiveness, and minimal cognitive load for visually impaired users. The "Walk Safe" and OCR functionalities are activated automatically based on real-time context, eliminating the need for continuous user input or manual toggling. The Vuzix Blade 2's built-in mono speaker delivers clear auditory alerts, and the Azure Cognitive Speech Service is used to generate natural-sounding speech for text output. During internal evaluations, the speech synthesis was found to be intelligible in quiet and moderately noisy environments. However, clarity may degrade in outdoor settings with high ambient noise. To mitigate this, future enhancements will consider adaptive volume control, ambient noise detection, and bone-conduction audio output to ensure clarity without isolating users from environmental sounds.

Although no formal user trials were conducted in this phase, the interface was iteratively tested in simulated user scenarios, ensuring the system responded consistently and rapidly to visual inputs. The UI architecture avoids menu hierarchies or gestures, relying instead on passive activation and immediate feedback—an approach intended to reduce interaction complexity and enhance user trust. A future usability study involving visually impaired participants is planned to empirically assess interaction intuitiveness, ease of learning, and overall user satisfaction.

6 Conclusion

This study introduces a novel system design that integrates object detection, OCR, and real-time feedback to provide a comprehensive assistive technology solution for visually impaired individuals. Leveraging Vuzix Blade 2 smart glasses and the robust YOLOv8 architecture, the system demonstrates significant advancements over existing assistive technologies by combining precision, scalability, and real-time usability. The critical comparative analysis of YOLOv8-N, YOLOv8-S, and YOLOv8-M highlights the trade-offs between computational efficiency and accuracy, with YOLOv8-S emerging as the most practical model for deployment on resource-constrained devices.

The innovative pipeline presented in this study not only addresses the limitations of prior solutions but also establishes a scalable framework for developers and researchers. By integrating cloud-based services with wearable devices, the system achieves a balance between local processing and advanced computational tasks, ensuring adaptability to various environments. The inclusion of OCR expands the system's functionality, enabling users to access textual information in real-time, a critical feature for environments such as university campuses. This research highlights the transformative potential of combining AR, AI, and computer vision in assistive technologies. The results, which demonstrate an overall mAP of 87.7% and near-perfect performance for high-frequency classes, validate the system's effectiveness and set a benchmark for future developments. Despite challenges with imbalanced datasets and low-frequency object detection, this study lays the foundation for inclusive and scalable navigation systems. Future enhancements, such as scene explanation, point-to-point navigation, and voice command integration, could further enhance the system's utility and impact, paving the way for broader adoption in diverse settings.

7 Limitations and future studies

Despite the advancements presented in this study, several limitations exist that provide opportunities for future development. One of the limitations is the system's reliance on predefined datasets, and the computational constraints of wearable devices, such as Vuzix Blade 2 smart glasses, may limit the scalability and performance in dynamic or unstructured environments. The dependency on a stable internet connection for real-time OCR and AI services also restricts usability in areas with poor connectivity. Another limitation is the current lack of point-to-point navigation, which would enable users to identify and move toward specific destinations within mapped areas. Incorporating this feature could greatly enhance the system's utility in complex settings, such as university campuses. Moreover, the application's interface requires further optimisation to allow hands-free operation through voice commands, enabling users to activate features like obstacle detection or OCR seamlessly. Another limitation worth mentioning is that the system demonstrates reduced detection performance under extreme lighting conditions, such as direct sunlight or low-light environments. While the model was trained on a diverse dataset to account for lighting variability, edge cases involving glare, reflections, or significant shadowing occasionally degraded the object detection confidence. Future iterations will explore adaptive exposure control and integration of infrared or thermal imaging to improve robustness in variable lighting. Battery constraints of the Vuzix Blade 2 present a practical limitation. With continuous use of camera-based processing and wireless communication, the average operational time was limited to approximately 2–3 h. This restricts extended usage in real-world applications. Potential mitigations include incorporating edge AI optimisation (e.g., quantised models), on-device inference scheduling, or external battery attachments to extend usability.

Additionally, although Azure Cognitive Speech Services deliver relatively natural text-to-speech output, speech clarity can be affected in noisy outdoor environments. Integration with noise-adaptive audio processing, bone-conduction output, or optional haptic feedback is a potential enhancement under consideration. Lastly, the current system lacks advanced scene understanding and dynamic path planning, which are critical for complete autonomous mobility. Future research will incorporate semantic segmentation, context-aware scene analysis, and conversational AI modules to provide a more intuitive and interactive assistive experience. A structured usability study will also be conducted to evaluate the system's acceptability, accessibility, and long-term impact in real-world contexts.

Future work should also address the need for broader situational awareness by integrating scene explanation capabilities, providing users with high-level descriptions of their surroundings. Expanding intelligent assistance features like conversational AI and context-aware prompts could make the system more adaptive to individual user needs, improving overall interactivity and accessibility. By addressing these limitations, future iterations of the system could achieve enhanced performance, usability, and inclusivity, making assistive technologies more accessible across diverse settings. Finally, it is important to note that this study did not involve human participants during the system evaluation phase. Therefore, it is planned to conduct a longitudinal user study with visually impaired participants. As such, usability, cognitive load, and accessibility aspects remain untested. A structured usability study with visually impaired participants is planned for future research. This two-stage validation—technical and user-oriented—ensures that the system is not only functional but also practically implementable. This study will follow ethical review protocols and assess user acceptance, system learnability, task completion accuracy, and subjective satisfaction. Including real-world feedback will be essential for validating the system's practical applicability and ensuring that it aligns with the daily mobility and information needs of end-users.

Appendix

See Figures 10, 11, 12, 13, 14, 15.

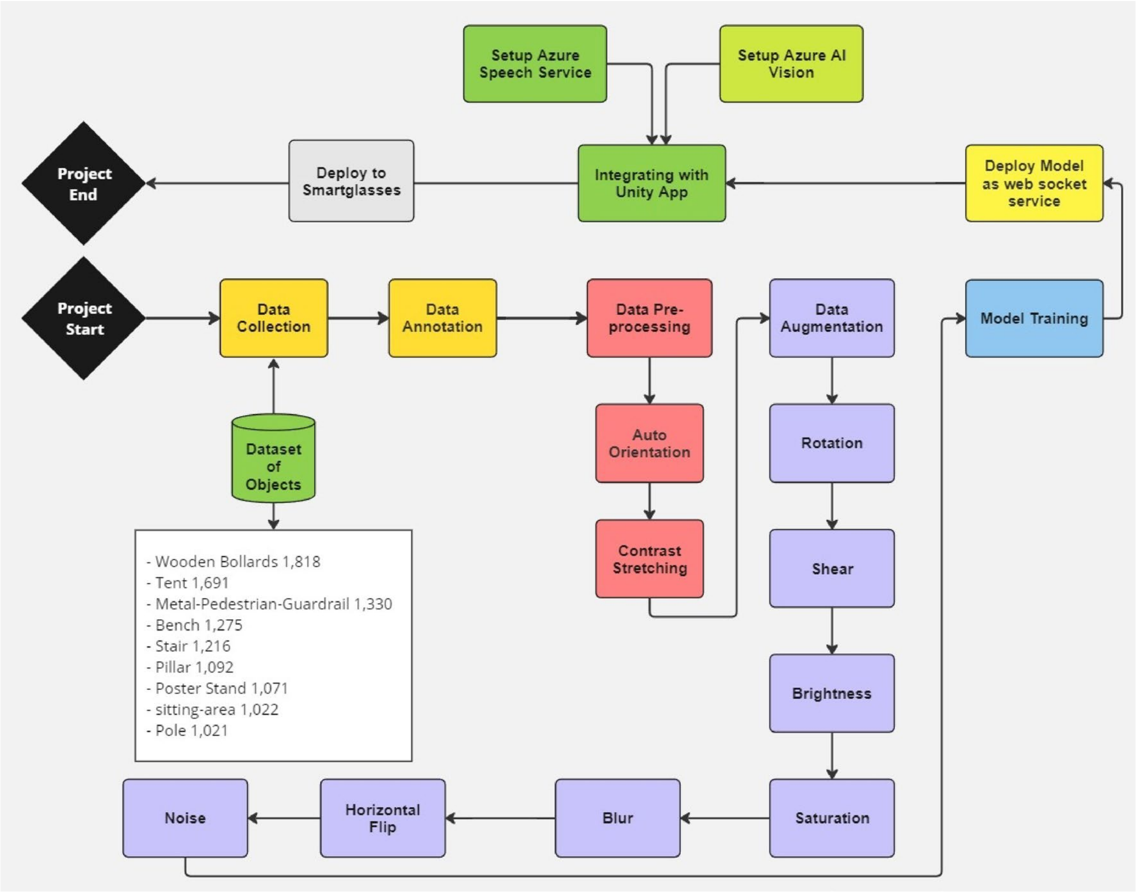


Fig. 10 System Design flow

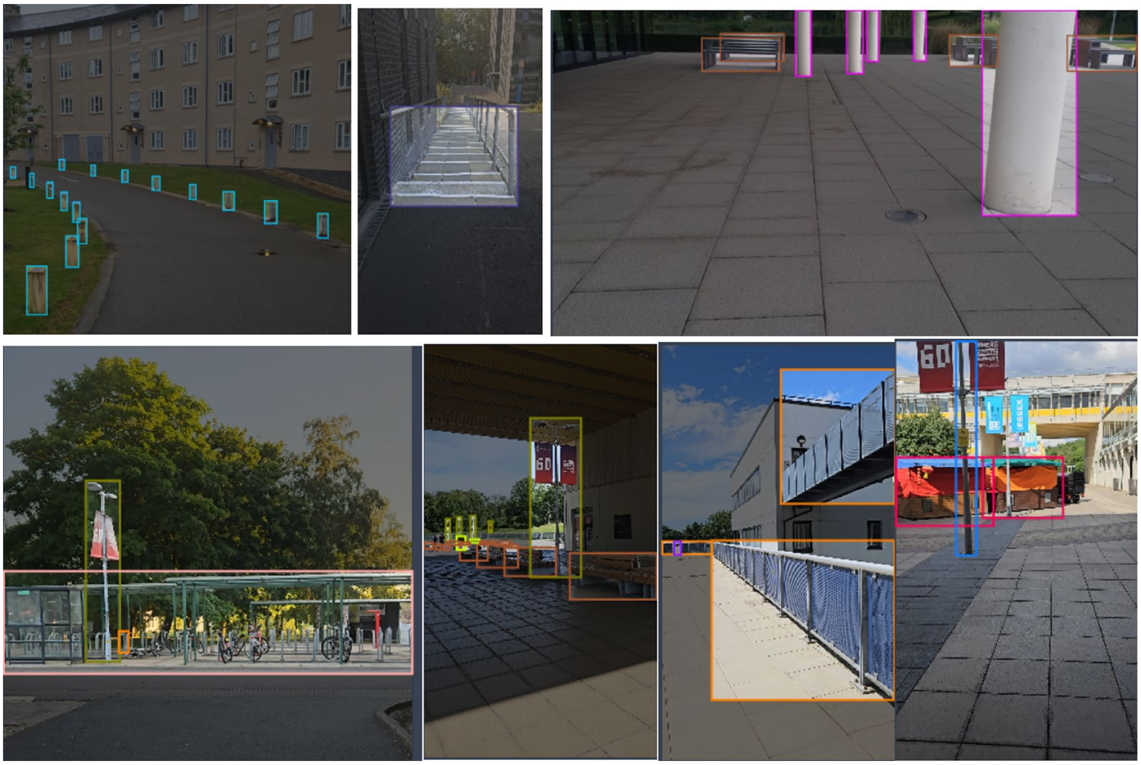


Fig. 11 Examples of the annotated objects taken at the university

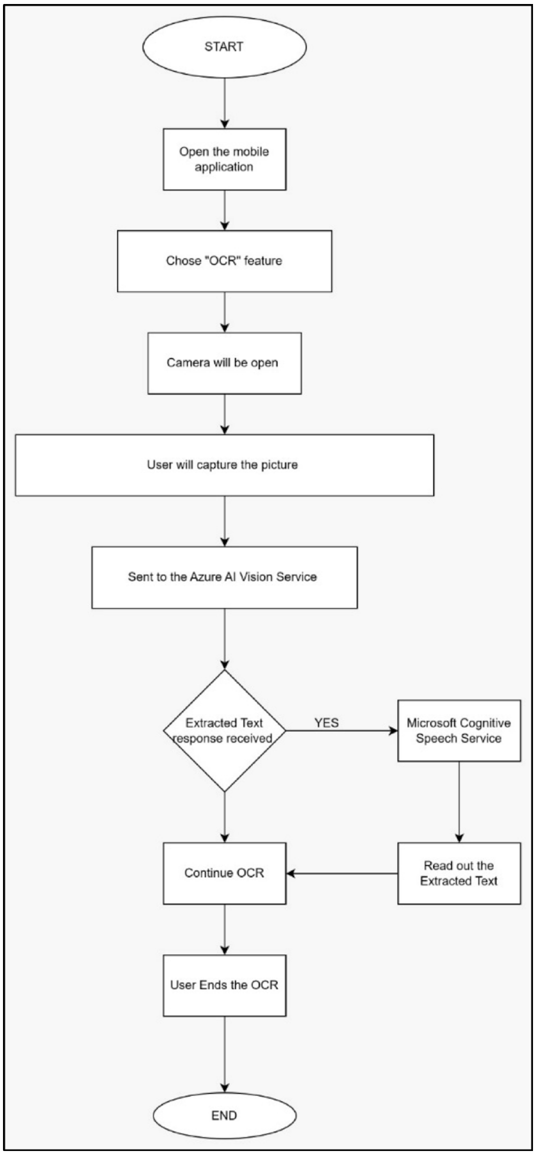

START
Open the mobile application
Chose "OCR" feature
Camera will be open
User will capture the picture
Sent to the Azure AI Vision Service
Extracted Text response received
YES
Microsoft Cognitive Speech Service
Read out the Extracted Text
Continue OCR
User Ends the OCR
END

Fig. 12 User Journey

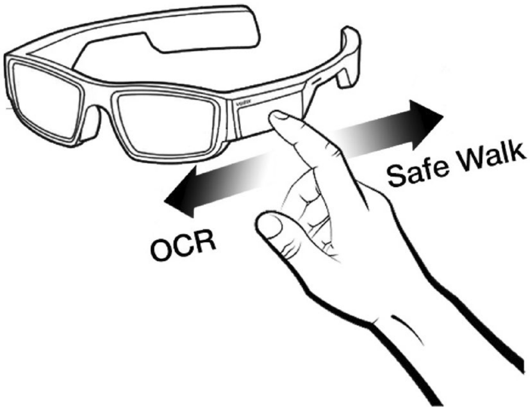


Fig. 13 System Gesture UI with Vuzix Blade 2

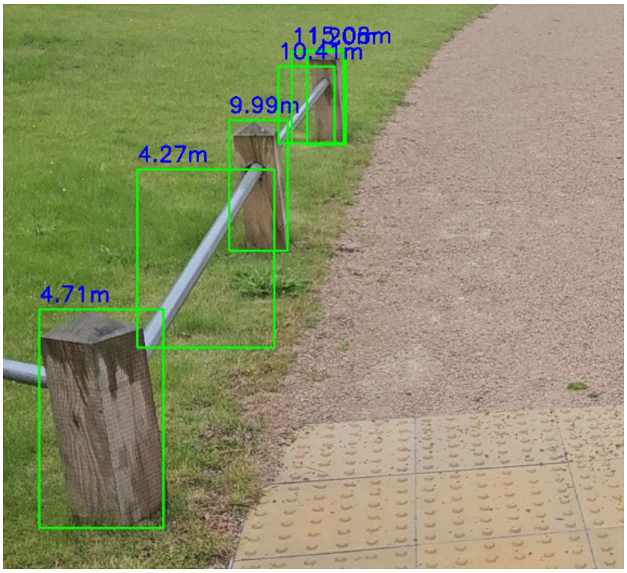


Fig. 14 Distance Calculation

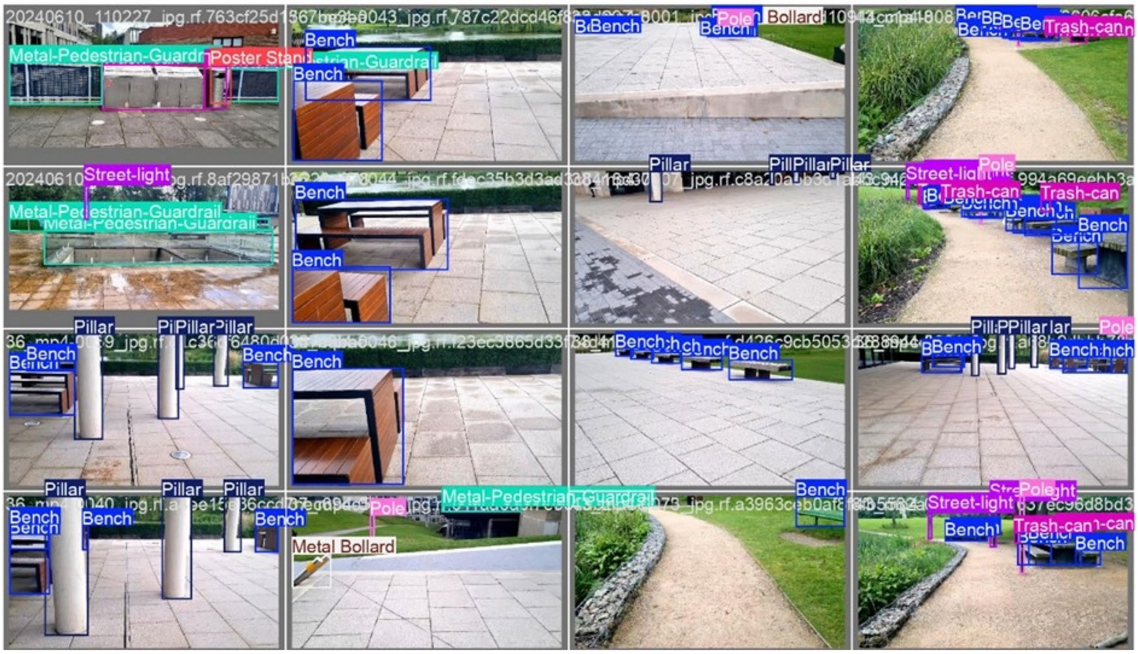


Fig. 15 System identifications by YOlOv8-M

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00530-025-02139-z.

Author contributions R.H. and P.K. conceptualised the study, carried out the literature review, and P.K. led the development of the object detection and OCR modules. R.H. contributed to the system architecture, oversaw the experimental design, and supported the model training and evaluation. Both authors jointly developed the methodology and implementation pipeline. P.K. and R.H. wrote the manuscript and prepared the figures and tables. All authors reviewed and approved the final manuscript.

Funding No funding was received for this study.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Ali, M., Zhang, Z.: The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection. Computers 2024a, 13, 336. In. (2024)
2. Ali, M.L., Zhang, Z.: The YOLO framework: A comprehensive review of evolution, applications, and benchmarks in object detection. Computers. **13**(12), 336 (2024b)
3. Alter, J., Xue, J., Dimnaku, A., Smirni, E.: *SSD failures in the field: symptoms, causes, and prediction models.* Paper presented at the Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. (2019)
4. Amore, F., Silvestri, V., Guidobaldi, M., Sulfaro, M., Piscopo, P., Turco, S., Rizzo, S.: Efficacy and patients' satisfaction with the ORCAM MyEye device among visually impaired people: A multicenter study. J. Med. Syst. **47**(1), 11 (2023)
5. Aziz, L., Salam, M.S.B.H., Sheikh, U.U., Ayub, S.: Exploring deep learning-based architecture, strategies, applications and current trends in generic object detection: A comprehensive review. IEEE Access. **8**, 170461–170495 (2020)
6. Benkirat, I.: Design and Implementation of a Real-Time Object Detection and Understanding System Using Deep Neural Network To Assist the Visually Impaired Persons. running on Raspberry Pi (2023)
7. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: *Simple online and realtime tracking.* Paper presented at the 2016 IEEE international conference on image processing (ICIP). (2016)
8. Bhuyan, M.K.: Computer Vision and Image Processing: Fundamentals and Applications. CRC (2019)
9. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*. (2020)
10. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: *End-to-end object detection with transformers.* Paper presented at the European conference on computer vision. (2020)

11. Cassinelli, A., Sampaio, E., Joffily, S., Lima, H., Gusmão, B.: Do blind people move more confidently with the tactile radar? Technol. Disabil. **26**(2–3), 161–170 (2014)
12. Casas, E., Ramos, L., Bendek, E., Rivas-Echeverría, F.: Assessing the effectiveness of YOLO architectures for smoke and wildfire detection. IEEE Access. (2023)
13. Chen, W., Luo, J., Zhang, F., Tian, Z.: A review of object detection: Datasets, performance evaluation, architecture, applications and current trends. Multimedia Tools Appl., 1–59. (2024)
14. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: *Yolo-world: Real-time open-vocabulary object detection.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2024)
15. Dalal, N., Triggs, B.: *Histograms of oriented gradients for human detection.* Paper presented at the 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). (2005)
16. Doe, J., & Smith, J.: Enhancing Out-of-Model Scope Detection with TRust Your GENerator (TRYGEN). http://www.naisjournal.com/static/upload/file/20250326/1742974333101187.pdf. (2021)
17. Eigen, D., Puhrsch, C., & Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems, 27 (2014)
18. Elavarasu, M., Govindaraju, K.: Unveiling the advancements: YOLOv7 vs YOLOv8 in pulmonary carcinoma detection. J. Rob. Control (JRC). **5**(2), 459–470 (2024)
19. Everding, L., Walger, L., Ghaderi, V.S., Conradt, J.: *A mobility device for the blind with improved vertical resolution using dynamic vision sensors.* Paper presented at the 2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom). (2016)
20. Fani Sani, M., Vazifehdoostirani, M., Park, G., Pegoraro, M., van Zelst, S.J., van der Aalst, W.M.: *Event log sampling for predictive monitoring.* Paper presented at the International Conference on Process Mining. (2021)
21. Girshick, R.: Fast r-cnn. *arXiv preprint arXiv:1504.08083*. (2015)
22. Gopalakrishnan, S., Kartha, A., Schuchard, R., Fletcher, D.: Comparison of visual function analysis of people with low vision using three different models of augmented reality devices. *medRxiv*, 2024.2009. 2011.24313484. (2024)
23. Granquist, C., Sun, S.Y., Montezuma, S.R., Tran, T.M., Gage, R., Legge, G.E.: Evaluation and comparison of artificial intelligence vision Aids: Orcam Myeye 1 and seeing Ai. J. Visual Impairment Blindness. **115**(4), 277–285 (2021)
24. Hartley, R., & Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2004)
25. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015)
26. He, J., Song, X., Su, Y., Xiao, Z.: A smart obstacle avoiding technology based on depth camera for blind and visually impaired people. CCF Trans. Pervasive Comput. Interact. **5**(4), 382–395 (2023)
27. He, C., Saha, P.: Investigating YOLO models towards outdoor obstacle detection for visually impaired people. ArXiv Preprint. (2023). arXiv:2312.07571
28. Henrique, V.: Image Enhancement: A Brief Introduction to Image Enhancement. (2023). Retrieved from https://medium.com/@henriquevedoveli/image-enhancement-4e18c1767c7
29. Hussain, M.: YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. Machines. **11**(7), 677 (2023)
30. Hussain, M.: Yolov5, yolov8 and yolov10: The go-to detectors for real-time vision. *arXiv preprint arXiv:2407.02988*. (2024)
31. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*. (2017)
32. Imhoff, J.: The Lancet Global Health: Vision loss could be treated in one billion people worldwide, unlocking human potential and accelerating global development. (2021)., 6 Jan Retrieved from https://www.michiganmedicine.org/news-release/lancet-global-health-vision-loss-could-be-treated-one-billion-people-worldwide
33. Jafri, R., Campos, R.L., Ali, S.A., Arabnia, H.R.: Visual and infrared sensor data-based obstacle detection for the visually impaired using the Google project Tango tablet development kit and the unity engine. IEEE Access. **6**, 443–454 (2017)
34. Jakob: Vuzix Blade 2 Review. (2023). Retrieved from https://vrx.vr-expert.com/vuzix-blade-2-review-vrx-by-vr-expert/
35. Jia, F., Afaq, M., Ripka, B., Huda, Q., Ahmad, R.: Vision-and Lidar-Based autonomous Docking and recharging of a mobile robot for machine tending in autonomous manufacturing environments. Appl. Sci. **13**(19), 10675 (2023)
36. Jocher, G., Stoken, A., Chaurasia, A., Borovec, J., Kwon, Y., Michael, K., Hogan, A.: ultralytics/yolov5: v6.0-YOLOv5n'Nano'models, Roboflow integration, TensorFlow export, OpenCV DNN support. *Zenodo*. (2022)
37. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural. Inf. Process. Syst., **25**. (2012)
38. Li, B., Munoz, J.P., Rong, X., Chen, Q., Xiao, J., Tian, Y., Yousuf, M.: Vision-based mobile indoor assistive navigation aid for blind people. IEEE Trans. Mob. Comput. **18**(3), 702–714 (2018)
39. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: *Feature pyramid networks for object detection.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition. (2017)
40. Lorenzini, M.-C., Jarry, J., Wittich, W.: The impact of using eSight eyewear on functional vision and oculo-motor control in low vision patients. Investig. Ophthalmol. Vis. Sci. **58**(8), 3267–3267 (2017)
41. Lu, Y.-F., Yu, Q., Gao, J.-W., Li, Y., Zou, J.-C., Qiao, H.: Cross stage partial connections based weighted Bi-directional feature pyramid and enhanced Spatial transformation network for robust object detection. Neurocomputing. **513**, 70–82 (2022)
42. Metz, R.: Augmented-Reality Glasses Could Help Legally Blind Navigate. (2015). Retrieved from https://www.technologyreview.com/2015/06/15/72902/augmented-reality-glasses-could-help-legally-blind-navigate/
43. Miura, T.: Narrative review of assistive technologies and sensory substitution in people with visual and hearing impairment. Psychologia. **65**(1), 70–99 (2023)
44. Mone, G.: Feeling sounds, hearing sights. Commun. ACM. **61**(1), 15–17 (2017)
45. Pasqualotto, A., Esenkaya, T.: Sensory substitution: The Spatial updating of auditory scenes mimics The Spatial updating of visual scenes. Front. Behav. Neurosci. **10**, 79 (2016)
46. Pii, J.: Vuzix Blade 2 Review. (2023). Retrieved from https://vrx.vr-expert.com/vuzix-blade-2-review-vrx-by-vr-expert/
47. Real, S., Araujo, A.: Navigation systems for the blind and visually impaired: Past work, challenges, and open problems. Sensors. **19**(15), 3404 (2019)
48. Redmon, J.: *You only look once: Unified, real-time object detection.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition. (2016)
49. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: *You only look once: Unified, real-time object detection.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition. (2016)
50. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2016)

51. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural. Inf. Process. Syst., 28. (2015)
52. Rodrigues, J.J., Aguiar, P.M., Xavier, J.M.: *Ansig—an analytic signature for permutation-invariant two-dimensional shape representation.* Paper presented at the 2008 IEEE Conference on Computer Vision and Pattern Recognition. (2008)
53. Saputra, M.R.U., Santosa, P.I.: *Obstacle avoidance for visually impaired using auto-adaptive thresholding on Kinect's depth image.* Paper presented at the 2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops. (2014)
54. Smith, R.: An overview of the TesseractOCR engine. In Ninth international conference on document analysis and recognition (ICDAR2007) (Vol. 2, pp. 629-633). IEEE (2007)
55. Szeliski, R.: Image processing. In Computer Vision: Algorithms and Applications (pp. 87-180). London: Springer London (2010)
56. Tan, M., Pang, R., Le, Q.V.: *Efficientdet: Scalable and efficient object detection.* Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020)
57. Terven, J., Córdova-Esparza, D.-M., Romero-González, J.-A.: A comprehensive review of Yolo architectures in computer vision: From Yolov1 to Yolov8 and Yolo-nas. Mach. Learn. Knowl. Extr. **5**(4), 1680–1716 (2023)
58. Trigka, M., Dritsas, E.: A comprehensive survey of machine learning techniques and models for object detection. Sensors. **25**(1), 214 (2025)
59. Ultralytics: Ultralytics YOLO Docs. (2025). Retrieved from https://docs.ultralytics.com/tasks/detect/
60. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vision. **57**, 137–154 (2004)
61. Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y.M.: *Scaled-yolov4: Scaling cross stage partial network.* Paper presented at the Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. (2021)
62. Webson, A.: Eye health and the decade of action for the sustainable development goals. Lancet Global Health. **9**(4), e383–e384 (2021)
63. Wojke, N., Bewley, A., Paulus, D.: *Simple online and realtime tracking with a deep association metric.* Paper presented at the 2017 IEEE international conference on image processing (ICIP). (2017)
64. Wang, C.-Y., Liao, H.-Y.M.: YOLOv1 to YOLOv10: The fastest and most accurate real-time object detection systems. APSIPA Trans. Signal. Inform. Process., **13**(1). (2024)
65. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., Ding, G.: Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*. (2024)
66. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*. (2020)
67. Zerroug, A., Cassinelli, A., Ishikawa, M.: Virtual haptic radar. In *ACM SIGGRAPH ASIA 2009 Sketches* (pp. 1–1). (2009)
68. Zhang, H., Ye, C.: An indoor wayfinding system based on geometric features aided graph SLAM for the visually impaired. IEEE Trans. Neural Syst. Rehabil. Eng. **25**(9), 1592–1604 (2017)
69. Zhang, C., Ding, W., Peng, G., Fu, F., & Wang, W.: Street view text recognition with deep learning for urban scene understanding in intelligent transportation systems. IEEE Trans. Int. Trans. Syst. **22**(7), 4727-4743 (2020)
70. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Chen, J.: *Detrs beat yolos on real-time object detection.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2024)
71. Zhou, J., Zhao, W., Guo, L., Xu, X., & Xie, G.: Real time detection of surface defects with inception-based Mobile Net-SSD detection network. InInternational Conference on Brain Inspired Cognitive Systems (pp. 510-519). Cham: Springer International Publishing (2019)
72. Zöllner, M., Huber, S., Jetter, H.-C., Reiterer, H.: *NAVI–a proof-of-concept of a mobile navigational aid for visually impaired based on the microsoft kinect.* Paper presented at the Human-Computer Interaction–INTERACT 2011: 13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5–9, 2011, Proceedings, Part IV 13. (2011)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.