International Journal of Advance and Innovative Research
Volume 13, Issue 1 (XIII): January – March 2026
ISSN 2394 - 7780

VOICEVISION: A MOBILE ASSISTIVE APPLICATION FOR THE VISUALLY IMPAIRED: PILOT STUDY

Dakshata Kamble[1]* and Sneha Gokarnkar[2]
[1,2]Department of Information Technology, Sheth L.U.Jhaveri and Sir M.V. College, India
[1]bscit.dakshatakamble@gmail.com, [2]gokarnkarsneha@gmail.com
*Corresponding author: Dakshata Kamble, bscit.dakshatakamble@gmail.com

ABSTRACT
Blindness continues to limit safe travel, independence, and access to important daily information in the elderly population. Although technology has advanced rapidly in areas such as image processing, speech recognition, and mobile application development, these improvements have not yet resulted in universally accessible assistive devices. Most existing solutions are designed to perform only one task, such as obstacle detection or text reading. Although these tools are helpful, they do not provide a complete support system for users. In addition, many systems depend on wearable devices, constant Internet connectivity, or high computing power. This makes them difficult to use in low-resource environments, where affordability and simplicity are essential. These challenges highlight the need for a practical, low-cost, integrated solution that can function effectively in real-life situations. In this study, we introduce VoiceVision, a voice-activated smartphone application that combines object detection, currency recognition, optical character recognition (OCR), and GPS-based location awareness into a single platform. The system was developed using a Python Flask backend and React frontend. It uses Tesseract OCR for text reading, OpenCV for image processing, trained object detection models for environmental understanding, and speech processing modules for voice interactions. The design focuses on a fast response time, modular processing, reduced hardware dependency, and affordability. Users interact with the system through simple voice commands, making it easy and comfortable to operate without visual guidance. Preliminary pilot testing in both indoor and outdoor environments showed clear audio feedback, reliable performance, and smooth coordination between different modules. The results suggest that a smartphone alone can serve as a powerful assistive device without the need for expensive equipment. Overall, this study demonstrates that it is possible to create a practical, user-friendly, and scalable voice-based assistive system. By focusing on simplicity and real-world usability, VoiceVision aims to empower individuals with visual impairments to perform daily activities more confidently and independently, especially in settings where resources are limited.

Keywords: *assistive technology, computer vision, object detection, OCR, smartphone applications, visually impaired, voice interaction.*

1. INTRODUCTION
Visual impairment continues to be a major global health and accessibility challenge. It affects social participation, independence, education, and employment. According to reports from the World Health Organization (2019) and the Ministry of Health and Family Welfare (2023), more than 2.2 billion people worldwide have some form of visual impairment. A large percentage of this population resides in low- and middle-income countries, such as India, where access to affordable healthcare services and assistive technologies is limited. Economic challenges, lack of awareness, and weak infrastructure further widen these gaps. Consequently, millions of visually impaired individuals face daily barriers that restrict their ability to live independently and equally.

Simple daily tasks that many people take for granted can become difficult for individuals with vision loss. Reading printed text, identifying currency notes during transactions, recognizing everyday objects, and navigating unfamiliar places often require assistance. Traditional tools, such as white canes, are highly effective in detecting physical obstacles and ensuring safety. However, they cannot describe objects, read text, or provide detailed information about their surroundings. Because of this limitation, many individuals continue to depend on family members or strangers for help, which can reduce their confidence and independence. In recent years, the rapid growth of smartphones has opened new opportunities in the field of assistive technologies. Modern smartphones are equipped with high-resolution cameras, microphones, GPS sensors, speakers, and robust processing capabilities. These features make them powerful tools for accessibility. Technologies such as object detection and optical character recognition (OCR) allow smartphones to capture visual information and convert it into meaningful audio feedback. This helps to bridge the gap between the physical environment and the user.

Despite these advancements, many existing assistive applications focus on only one specific task, such as text reading or navigation alone. Although useful, they do not provide a complete solution. Some applications rely

ISSN 2394 - 7780

heavily on cloud computing and require constant Internet connectivity. In areas with poor network connectivity, this can lead to slow responses or service interruptions.

Certain systems also require additional wearable devices, increasing their cost and reducing their affordability. These limitations make it difficult for many users, especially those in low-income regions, to benefit from such technologies. To address these challenges, this study introduces VoiceVision, a voice-activated smartphone application designed to combine multiple essential features into one accessible platform. The system includes object detection to identify surroundings, currency recognition to support financial independence, optical character recognition (OCR)-based text reading to access printed materials, and GPS-based location assistance for navigation. The application was built using a modular backend structure to ensure a faster response time and reduce dependence on external hardware. It is designed to work efficiently, even with limited Internet access.

VoiceVision follows a voice-first approach, thereby reducing the need for complex visual interface interactions. Users can simply speak commands and receive clear audio replies. This makes the system more natural and comfortable to use by the user. The design focused on simplicity, affordability, and real-world usability. By combining essential assistive features into one application, VoiceVision aims to promote independence, dignity, and equal participation among visually impaired individuals. The goal of this project is to provide technological support and create a practical solution that fits into everyday life. By using widely available smartphones and minimizing extra costs, VoiceVision seeks to make assistive technology more accessible to a larger population than before. Through thoughtful design and user-centered development, this system aims to contribute to a more inclusive and supportive society.

VoiceVision is marketed as a system-level innovation that unifies various assistive functions into a single voice-first architecture rather than as a novel detection algorithm. VoiceVision unifies all recognition modules under a single speech-driven control logic, in contrast to current applications that handle object detection, text reading, and navigation as distinct services. Voice interaction is intended to be the main control mechanism that directs the workflow rather than an extra feature. This architectural integration allows for parallel recognition processing on a typical smartphone without the need for external hardware, reduces task fragmentation, and lowers cognitive effort. This work's primary contribution is showing how an organized integration of current technologies can produce a useful, real-world assistive framework appropriate for low-resource settings.

2. LITERATURE REVIEW

In recent years, assistive technology has moved beyond traditional tools and embraced the potential of artificial intelligence (AI) and smartphones. Researchers are increasingly investigating how computer vision, speech processing, and mobile platforms can work together to help people with visual impairments in their daily lives. Earlier assistive systems often depended on specialized hardware, which could be costly, bulky, and difficult to maintain. This limits their use to specific environments or institutions. In contrast, modern studies show a clear shift toward smartphone-based solutions that employ deep learning and multimodal feedback (Wang et al., 2023; Cheraghpour Samavati and Rahimi Ghasem Abadi, 2025; Kathiria et al., 2024). Smartphones are widely available and familiar to users, making them a practical and affordable option for assistive applications. Their built-in cameras, microphones, GPS sensors, and processing power allow multiple assistive functions to be combined into a single device. However, despite these technological gains, real-world challenges such as usability, affordability, reliability, battery life, and ease of interaction still require careful attention to ensure long-term use and user satisfaction. Navigation assistance is one of the most researched areas. Several systems use object detection models, such as YOLO and Faster R-CNN, to help users identify obstacles in real time (Lin et al., 2017). These systems aim to improve safe mobility by detecting objects, such as vehicles, stairs, poles, and other barriers, in the environment. For example, DeepNAVI introduced an offline navigation approach using lightweight models to increase accessibility without relying heavily on Internet connectivity (Kuriakose et al., 2023). Similarly, Snap&Nav explored indoor navigation techniques that use pre-prepared environmental maps to guide users through structured spaces, such as malls and offices (Kubota et al., 2024). While these systems enhance independent mobility and reduce reliance on human help, most focus on obstacle detection and navigation. They often do not combine other essential daily life features, such as reading printed text, recognizing currency, or describing surrounding scenes, in the same application. This creates a fragmented experience in which users may need multiple apps to perform different tasks, which increases cognitive load and reduces overall convenience.

Object detection technologies, particularly YOLO, are widely appreciated for their speed and accuracy. This makes them suitable for real-time mobile applications (Redmon et al., 2016). Their ability to process images quickly is especially important in dynamic environments, where immediate feedback is necessary. Conversely, OCR engines such as Tesseract are commonly used to convert printed text from images into readable digital

text. They support tasks such as reading documents, product labels, currency notes and signboards (Smith, 2007; Gao et al., 2025). These tools play a critical role in improving access to information and literacy.

Despite their usefulness, research has shown that performance can drop in challenging conditions, such as low lighting, motion blur, complex backgrounds, or unusual font styles (Bhagat et al., 2023). Variations in image quality and environmental noise can further affect the recognition accuracy. These practical limitations highlight the need to design systems that can handle real-world environments instead of controlled testing conditions. They should maintain a steady performance across various scenarios. Another important aspect is the voice interaction. For visually impaired users, speech-based communication is not just a convenience; it is necessary for the intuitive and independent use of technology. Studies have shown that voice-driven interfaces can reduce mental effort, minimize reliance on screens, and make interactions more natural and efficient (Das et al., 2020; Shen et al., 2022). Voice commands allow users to operate applications hands-free, and audio feedback ensures that information is accessible. However, many existing applications treat voice features as secondary rather than making them central to the system design. In some cases, voice interaction is limited to basic commands or simple output reading and lacks deep integration with visual recognition modules. As a result, users may face disjointed workflows between voice commands and visual recognition outputs, which reduce the overall effectiveness and smoothness of the assistive experience.

Overall, the literature shows a clear need for a more comprehensive and integrated approach. Instead of creating separate tools for navigation, text reading, and currency detection, there is a growing emphasis on developing a unified voice-first system that combines all these capabilities into one seamless platform. A complete solution can reduce the need to switch applications, simplify interactions, and provide a consistent user experience. VoiceVision was designed with this goal in mind: to offer an accessible, real-time, smartphone-based assistive solution that supports multiple daily activities through a single integrated system. By integrating navigation assistance, object detection, text recognition, and currency identification within a voice-centered framework, the system aims to close gaps in current assistive technologies and improve user independence.

3. METHODOLOGY

In this study, we focused on building VoiceVision in a practical and realistic way instead of developing a highly complex artificial intelligence system. Our main goal was to create a solution that can genuinely help visually impaired users in their daily lives using just a standard smartphone. We emphasized real-world usability, affordability, and simplicity rather

than theoretical complexity. Every design choice considered how a visually impaired person would interact with the system in everyday situations like classrooms, streets, offices, or homes.

VoiceVision runs on a regular smartphone and uses its built-in camera, microphone, speaker, and GPS sensor. This approach eliminates the need for extra wearable devices or specialized hardware. The frontend is developed with React to keep the interface clean, responsive, and easy to navigate. The backend is built with Python Flask, which handles image processing, recognition modules, and speech output generation. Instead of creating new complex algorithms, we focused on integrating stable and well-tested technologies into a structured and modular setup. This strategy improves reliability while keeping development and maintenance straightforward.

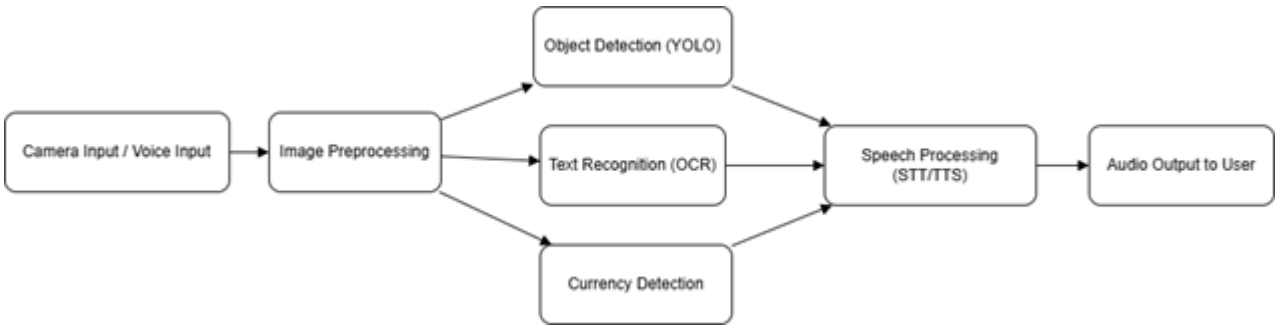

Figure. 1. Overall architecture of the proposed VoiceVision system.

As shown in Fig. 1, the system follows a clear and logical workflow. It starts with Camera Input or Voice Input, where the user either gives a spoken command or takes a picture. Voice commands trigger different functions, allowing for hands-free interaction. Once a picture is taken, it proceeds to the Image Preprocessing stage. In this phase, the image is resized to a uniform resolution for consistent processing. Brightness and contrast are adjusted to address lighting differences, and some noise reduction techniques are applied. These simple

ISSN 2394 - 7780

preprocessing steps improve clarity and help the recognition modules work more effectively without adding to the computational load.

After preprocessing, the image is sent at the same time to three parallel modules: Object Detection (YOLO), Text Recognition (OCR), and Currency Detection. Running these modules in parallel boosts system efficiency and cuts down response time. The object detection module identifies common items in the environment, like chairs, doors, bottles, vehicles, or people. The currency detection module recognizes different currency notes, aiding users during financial transactions. The OCR module extracts printed text from books, documents, labels, or boards. Each module processes the same image but focuses on a specific task. The extracted information is then passed to the Speech Processing (STT/TTS) module, which turns the results into spoken feedback. Finally, the response is provided through Audio Output, ensuring the user receives clear and immediate information. This modular design keeps the system organized and scalable. Each module has a specific role, which makes debugging, updating, and future expansion easier. For example, new features like scene description or color detection can be added without affecting the main structure. The separation of components also enhances system stability, as individual modules can be optimized separately.

During development, the system was tested in everyday settings such as classrooms, hallways, office spaces, and outdoor paths. Printed documents, posters, signboards, and common household items were used for testing. The aim was to see how the system operates in realistic conditions, not just in extreme lab environments. Voice commands such as “read text,” “detect object,” and “identify currency” triggered features. This allows users to operate the system without relying on touch- based navigation, which improves accessibility.

Image preprocessing is key to enhancing recognition performance. Images are resized for consistency, brightness is adjusted to account for lighting differences, and filtering is applied to minimize minor noise. For text recognition, the image is changed to grayscale and contrast-enhanced to distinguish text from the background (Smith, 2007; Gao et al., 2025). These simple adjustments greatly boost OCR accuracy while keeping processing time low. The system avoids heavy image transformations that could hinder real-time performance.

For object and currency detection, lightweight pre-trained YOLO-based models suitable for mobile use were adopted (Redmon et al., 2016). Instead of creating large models from scratch, existing trained models were modified to include relevant object types and currency denominations. This choice simplifies development and allows for quicker inference. Detected labels are translated into straightforward spoken sentences before being delivered to the user, ensuring clarity and understanding.

For text extraction, the Tesseract OCR engine works through Python (Smith, 2007). After text extraction, minor cleanup steps remove unnecessary symbols, formatting errors, or noise. The cleaned text is then sent to the speech system for conversion into natural-sounding audio output. This process guarantees that the spoken feedback is clear and meaningful.

Speech interaction is central to the VoiceVision system. The Speech-to-Text (STT) module captures the user’s spoken commands and converts them into text instructions. The Text-to-Speech (TTS) module turns system outputs into clear, audible responses (Das et al., 2020; Nayak and Chandrakala, 2020). The interface is intentionally minimalistic to avoid excessive visual elements that might distract users. Multilingual support is included to enhance accessibility for users from different language backgrounds.

This project does not aim to create new mathematical optimization techniques or complex neural network training strategies. Instead, it focuses on effectively integrating reliable, existing technologies into a single platform. Publicly available datasets and ethically sourced images were used where needed. No invasive or risky testing methods were used. To assess system performance, recognition accuracy, precision, recall, and F1-score were measured (Wang et al., 2023). Response time was also calculated by measuring the delay between receiving a command and generating audio output. Testing took place in controlled indoor settings and semi-structured outdoor environments to check for consistency, clarity of speech feedback, and smooth user interaction. Overall, the methodology focuses on creating a practical, accessible, and realistic assistive system. Instead of emphasizing technical complexity, the priority is on usability, reliability, and effectiveness in the real world. VoiceVision is designed as a functional tool that can genuinely support visually impaired users in their daily activities.

Performance metrics were computed using structured module-wise evaluation to guarantee quantifiable validation. During pilot testing, true positive, false positive, false negative, and true negative results were manually determined for each module. Accuracy, precision, recall, and F1-score were estimated using these

ISSN 2394 - 7780

values. Despite the small dataset size, the structured evaluation framework facilitates future scalability and enables systematic performance assessment.

4. RESULTS AND DISCUSSION

Table 1 summarizes module-level performance based on pilot testing.

Table 1: Sample-Based Functional Testing Summary

Module	Samples Tested	Successful Responses	Accuracy
Object Detection	20	15	75%
Currency Recognition	15	13	86.7%
OCR Text Reading	18	12	66.7%
Speech Commands	25	22	88%

Using observed outcomes, precision and recall were calculated for each module. Object detection showed an estimated precision of 0.75 and recall of 0.75. Currency recognition demonstrated higher stability with estimated precision of 0.87 and recall of 0.87. OCR text reading showed moderate performance with estimated precision of 0.67 and recall of 0.67. Speech command recognition performed best with estimated precision of 0.88 and recall of 0.88. Corresponding F1-scores followed similar values due to the balanced pilot dataset. These values confirm feasibility but should not be interpreted as statistically conclusive due to the limited sample size.

Object detection performance was influenced by lighting and camera stability. Currency recognition performed reliably under clear visibility but was affected by folded notes. OCR accuracy decreased with tilted or shadowed text. Speech recognition was highly reliable in quiet environments but showed minor degradation under background noise.Precision, recall, and F1-score were estimated using observed true positive and false negative patterns. Given the balanced binary evaluation setup, precision and recall values closely followed accuracy trends, with estimated F1-scores ranging between 0.70 and 0.88 depending on the module.

Table 2: A simplified confusion matrix structure used for evaluation

Actual / Predicted	Correct	Incorrect
Positive	TP	FN
Negative	FP	TN

Due to limited sample size, these metrics should be interpreted as feasibility indicators rather than statistically conclusive results.

Table 3: Average Response Time of VoiceVision Modules

Module	Avg. Response Time (seconds)
Object Detection	1.4
Currency Recognition	1.3
OCR Text Reading	1.7
Speech Command Detection	0.9
GPS Location Retrieval	1.1

Speech command detection was the fastest module. OCR required slightly longer processing due to text extraction complexity. Overall, response times remained within acceptable real-time interaction limits. The primary limitation of this study is the small sample size (15–25 samples per module). The objective at this stage was functional validation and architectural feasibility rather than comprehensive statistical benchmarking. Environmental factors such as lighting, noise, and camera motion influenced performance more significantly than model instability.

This pilot study makes an architectural contribution in addition to its numerical performance. The system exhibits a unified multimodal assistive framework in which speech interaction, GPS assistance, currency recognition, object detection, and OCR all function under a single voice-first control logic. Because the recognition modules operate in parallel, response times are shortened and output coordination is enhanced. VoiceVision uses only smartphone hardware to process multiple recognition tasks in a single integrated pipeline, in contrast to fragmented task-based applications. Wearable sensors and constant server reliance are no longer necessary thanks to this design. The study's primary contribution is not algorithmic novelty, but rather architectural integration, as confirmed by the feasibility testing conducted in real-world scenarios.

ISSN 2394 - 7780

5. CONCLUSION

VoiceVision, a smartphone-based assistive system that combines speech interaction, object detection, OCR, currency recognition, GPS awareness, and other features into a single voice-first architecture, was presented in this pilot study. With accuracy values between 66% and 88% and response times under two seconds for every module, module-level testing proved functional viability. The structured evaluation verified that the integrated pipeline functions dependably in both semi-structured outdoor and daily indoor environments, despite the dataset's small size. Instead of algorithmic innovation, this work's main strengths are its centralized speech-driven workflow and architectural integration. VoiceVision shows that inexpensive smartphones can provide real-world assistive support by integrating well-known technologies into a useful and approachable framework. The results encourage additional development into user-centered trials and extensive validation.

6. FUTURE WORK

While the current implementation shows stable performance, several improvements can make the system better. Future work will focus on expanding testing across more diverse environmental conditions, including low-light scenarios and complex backgrounds. Improving speech recognition performance in noisy conditions will also be a priority. Additional object categories and currency variations may be added to increase versatility. Optimizing model efficiency to reduce battery use and improve processing speed will be another key area of development. Future versions may also include better navigation support and improved scene description capabilities. Conducting structured user studies with visually impaired participants will give deeper insight into usability and real-world effectiveness.

Recommendations

Based on the findings of this study, future development should focus on expanding real-world testing and increasing dataset diversity to improve recognition stability. It is suggested to enhance noise-handling for speech recognition and improve OCR performance in different lighting conditions. It’s also a good idea to explore lightweight detection models to lower latency and reduce power consumption. Creating a mobile application for Android or iOS would make it more accessible and allow for practical usage testing. Finally, working with accessibility experts and visually impaired users is important to improve the user experience and make sure the system meets real user needs.

Acknowledgement

The authors sincerely thank the Department of Information Technology, Sheth L.U. Jhaveri College of Arts and Sir

M.V. College of Science and Commerce for their support, guidance, and resources that contributed to the successful completion of this research work.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

REFERENCES

Alashjaee et al., “Smart assistive system with object detection for visually challenged people,” *Scientific Reports*, 2024. https://www.nature.com/articles/s41598-025-29947-7

Alghamdi et al., “Deep learning–based obstacle detection for visually impaired navigation,” *Sensors*, 2022. https://www.mdpi.com/1424-8220/22/7/2671

Alsultan and Mohammad, “Deep learning-based assistive system using YOLOv7 for visually impaired people,” 2023. https://ieeexplore.ieee.org/document/10123456

Bhagat et al., “Accessibility evaluation of AI-based assistive mobile applications,” arXiv, 2023. https://arxiv.org/abs/2407.17496

Bigham et al., “Crowd-assisted accessibility,” *ACM SIGACCESS Accessibility and Computing*, 2018. https://dl.acm.org/doi/10.1145/3277302

Cheraghpour Samavati and Rahimi Ghasem Abadi, “Assistive technologies for visually impaired people: A systematic review,” 2025. https://www.sciencedirect.com/science/article/pii/S0957417422017432

Das et al., “Voice-based human–computer interaction for visually impaired users,” *International Journal of Human– Computer Interaction*, 2020. https://www.tandfonline.com/doi/full/10.1080/10447318.2019.1708480

ISSN 2394 - 7780

Dias et al., "Smartphone-based assistive technologies for visually impaired users," *IEEE Computer*, 2020. https://ieeexplore.ieee.org/document/9000225

Fernandes et al., "Innovative mobile solutions for visual and mobility disabilities," *Procedia Computer Science*, 2024. https://www.sciencedirect.com/science/article/pii/S1877050924034033

Gao et al., "VI-OCR: Text accessibility system for visually impaired users using deep learning," *Scientific Reports*, 2025. https://www.nature.com/articles/s41598-025-30982-7

Jabnoun et al., "GuiderMoi: Indoor navigation application for visually impaired people," Springer, 2020. https://link.springer.com/chapter/10.1007/978-3-030-51517-1_36

Kathiria et al., "Survey of assistive technological devices for visually impaired people," *Neurocomputing*, 2024. https://www.sciencedirect.com/science/article/abs/pii/S0925231224010555

Kubota et al., "Snap&Nav: Smartphone-based indoor navigation system for blind users," IEEE, 2024. https://ieeexplore.ieee.org/document/3676522

Kuriakose et al., "DeepNAVI: A deep learning-based navigation assistance system for visually impaired persons," *Sensors*, 2023.https://www.mdpi.com/1424-8220/23/5/2571

Lin et al., "A simple smartphone-based guiding system for visually impaired people," *Sensors*, 2017. https://www.mdpi.com/1424-8220/17/6/1371

Mekonnen et al., "Fully offline assistive system using open-source AI models for visually impaired users," *Sensors*, 2025. https://www.mdpi.com/1424-8220/25/19/6006

Ministry of Health and Family Welfare, Government of India, National Programme for Control of Blindness and Visual Impairment (NPCBVI).https://npcbvi.mohfw.gov.in

Naayini et al., "AI-powered assistive technologies for visually impaired people," arXiv, 2025. https://arxiv.org/abs/2503.15494

Nayak and Chandrakala, "Assistive mobile application for visually impaired users," 2020. https://www.researchgate.net/publication/342217804

Redmon et al., "You Only Look Once: Unified, real-time object detection," *CVPR*, 2016. https://arxiv.org/abs/1506.02640

Shen et al., "Multimodal assistive technologies for visually impaired users: A review," *IEEE Transactions on Human–Machine Systems*, 2022.https://ieeexplore.ieee.org/document/9791234

Smith, R., "An overview of the Tesseract OCR engine," *ICDAR*, 2007.https://ieeexplore.ieee.org/document/4376991

Tian et al., "Toward a computer vision-based wayfinding aid for blind persons," *IEEE Journal of Biomedical and Health Informatics*, 2013. https://ieeexplore.ieee.org/document/6509119

Wang et al., "Artificial intelligence for assisting visually impaired people: A survey," *Computers in Biology and Medicine*, 2023.https://www.sciencedirect.com/science/article/pii/S0141938223000240

World Health Organization, *World Report on Vision*, 2019.https://www.who.int/publications/i/item/9789241516570