IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, March 2026
Impact Factor: 8.2

SmartVision: Bringing Sight through Sound with Real-Time Guidance

Sanket Pawar[1], Ketan Suryavanshi[2], Siddhesh Gawade[3], Prof. Rashmi Mahajan[4]
Department of Artificial Intelligence & Machine Learning[1,2,3,4]
Shivajirao S. Jondhale College of Engineering, Dombivli (E), Maharashtra, India

Abstract: *Visually impaired individuals face systemic challenges in navigating complex environments and identifying everyday objects independently. [1] To address these accessibility barriers, this review paper evaluates SmartVision, a comprehensive, voice-activated assistive web application designed to significantly enhance user autonomy. The system integrates advanced computer vision models to provide real- time, auditory scene descriptions and general object detection, while introducing a novel, on-demand similarity matching process for user-specific personal items. [3] For seamless accessibility, SmartVision employs a highly secure, completely hands-free facial recognition login mechanism powered by DeepFace and Firebase, removing traditional input barriers. Furthermore, the platform supports independent mobility through an integrated GPS-based route navigation module and prioritizes user safety via an intelligent, voice-triggered emergency alert system.[4].*

Keywords: *SmartVision*

I. INTRODUCTION

Over the last decade, the rapid advancement of Artificial Intelligence (AI) and computer vision has paved the way for transformative applications in assistive technology [1]. For individuals with visual impairments, navigating complex environments and recognizing everyday objects consistently remains a fundamental challenge for independent living. While traditional mobility aids like white canes and guide dogs offer essential physical assistance, they are limited in their capacity to interpret detailed environmental contexts or read digital information [2]. As a result, researchers have increasingly focused on developing smart mobile systems that leverage deep learning to provide real-time auditory feedback [3].

Despite these technological strides, integrating these solutions into a cohesive, user-friendly platform that functions reliably in real-world scenarios continues to be a major obstacle [4]. Many existing applications suffer from high processing latency, complex manual interfaces, or require expensive specialized hardware, which restricts their overall accessibility. This highlights the urgent need for a more adaptable, lightweight, and perceptive human- machine interface for assistive use [5]. The SmartVision platform addresses these exact gaps by uniting real-time object detection, scene description, and GPS-assisted navigation into an accessible, browser- based environment [6]. Experimental observations indicate that SmartVision delivers accurate environmental interpretations with minimal delay, proving its effectiveness and viability for daily, real-world deployment [7].

By capturing live video streams through a standard smartphone or computer webcam, the system processes surrounding visual data dynamically and converts it into immediate auditory cues. Furthermore, to eliminate physical interaction barriers entirely, SmartVision introduces a completely hands-free facial recognition login mechanism, alongside a customized, on-demand personal object matching system that operates purely through intuitive voice commands [8]. Unlike conventional assistive applications that heavily tax local device hardware, this system optimizes its computational load by intelligently handling intensive tasks—such as spatial mapping and vector similarity searches—through highly efficient backend services [9]. This seamless integration of multimodal capabilities ultimately creates a robust framework that prioritizes the continuous safety, accessibility, and autonomy of visually impaired users [10].

DOI: 10.48175/IJARSCT-31609

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, March 2026 Impact Factor: 8.2

To ensure this high level of reliability and speed, the system leverages a robust cloud infrastructure that handles real-time data streaming without overwhelming the client device [11]. The inclusion of WebRTC protocols specifically prevents latency bottlenecks, facilitating instantaneous communication between the user's camera feed and the backend vision processing engines [12]. Beyond standard navigation and object recognition, SmartVision introduces a proactive safety layer designed explicitly to protect vulnerable users in unpredictable environments [13]. By integrating an intelligent, voice-activated emergency distress signaling mechanism that can optionally be triggered via hardware button presses, the platform ensures that help is immediately accessible during critical moments [14].The combination of these secure architectural choices and advanced cognitive services positions SmartVision as a highly effective, modern solution for dramatically enhancing the daily mobility and confidence of visually impaired[15].

II. REVIEW METHODOLOGY

In this work, we propose a comprehensive and modular methodology for the SmartVision system, designed to address the multifaceted challenges visually impaired individuals face during independent navigation [16]. Unlike traditional assistive devices that rely on single- sensor inputs, our methodology integrates multimodal perception, cloud-based deep learning inference, and intuitive conversational interfaces to create a holistic, real- time assistive environment. The architectural framework is structured into several interconnected pipelines, each optimized for low latency, high accuracy, and maximum accessibility.

Multimodal Perception and Data Acquisition The foundational layer of the SmartVision methodology involves continuous, real-time data acquisition through standard device cameras and microphones [17]. To ensure seamless user interaction, the system captures live WebRTC video streams and processes natural language voice commands simultaneously. This data is dynamically preprocessed on the client-side to normalize lighting, reduce background noise, and extract essential features before being securely transmitted to the backend inference engines. By intelligently balancing the computational load between edge- capture devices and cloud-processing servers, our methodology significantly reduces the hardware requirements for the end user [18].

Secure Hands-Free Authentication Pipeline To completely eliminate physical interaction barriers for visually impaired users, we implemented a robust, hands-free authentication mechanism utilizing the DeepFace framework [19]. Instead of relying on traditional email and password inputs, the system continuously scans the initial camera feed to extract facial embeddings. These encodings are then compared against securely stored profiles in a Firebase Firestore database using cosine similarity metrics. Upon a successful biometric match, the system automatically generates a custom session token, granting immediate, secure access to the core assistive dashboard without requiring a single screen tap.

Real-Time Detection and Scene Understanding At the core of the SmartVision methodology is the continuous, dynamic analysis of the user's physical surroundings. The system leverages optimized convolutional architectures to perform rapid object detection and generate detailed, contextual scene descriptions.When a video frame is captured, it is routed through the central computer vision pipeline, which identifies potential obstacles , human presence, and common environmental features [20].

These visual insights are instantly interpreted and converted into immediate auditory feedback through an integrated Text-to-Speech (TTS) module, minimizing the cognitive load on the user.

On-Demand Personal Object Similarity Search A key algorithmic novelty in our proposed methodology is the highly optimized retrieval system for identifying user-specific personal items. Recognizing that continuous vector similarity searches are computationally expensive and prone to latency, we designed a targeted, on-demand matching architecture [21]. The system allows users to store encoded representations of their unique personal objects (e.g., keys, wallets) in a secure vector database. When the user explicitly issues a voice command requesting to find an item, the platform cross-references the current camera frame against these stored embeddings using threshold- based similarity metrics, preserving system bandwidth while maximizing daily utility.

Context-Aware Navigation and Safety Protocols Beyond visual interpretation, our methodology inherently integrates spatial awareness and proactive safety protocols [16]. By integrating with external geolocation services like the Google

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, March 2026 Impact Factor: 8.2

Maps API, the system can provide dynamic, step-by-step auditory routing tailored for pedestrian transit. Furthermore, recognizing the vulnerabilities of visually impaired users in unfamiliar environments, we embedded a critical emergency distress framework. This safety layer can be triggered instantly via specific voice keywords or hardware button combinations, immediately broadcasting the user's exact geographical coordinates to predefined emergency contacts [17]. Through this multifaceted methodology, SmartVision operates not merely as an observational tool, but as a comprehensive lifeline for independent living.

III. LIMITATIONS OF EXISTING TECHNIQUES

While significant progress has been made in the development of computer vision applications for the visually impaired, existing assistive technologies still face several critical limitations that hinder their widespread adoption and daily usability [12]. These challenges can broadly be categorized into computational constraints, accessibility barriers in user interfaces, and the lack of personalized contextual awareness.

High Computational Overhead and Latency A primary limitation of modern deep learning-based assistive systems is their heavy reliance on complex neural network architectures, which demand substantial computational resources and high-end Graphical Processing Units (GPUs) [13]. When deployed on standard mobile devices, these models frequently suffer from high inference latency, leading to delayed auditory feedback. For a visually impaired user navigating a dynamic environment, even a processing delay of a few seconds can pose severe safety risks [14]. Furthermore, continuous real-time video processing rapidly drains device batteries and causes thermal throttling, making many existing applications impractical for extended, all- day use.

Accessibility Barriers in User Interfaces Paradoxically, many applications designed for the visually impaired still rely on traditional graphical user interfaces that require precise physical interactions [15]. Tasks such as creating an account, entering complex passwords, or navigating intricate menu structures to activate specific features present significant usability obstacles. Even systems that incorporate voice commands often require the user to manually tap a physical button to initiate listening mode, breaking the illusion of true hands-free autonomy [16]. These interaction bottlenecks negate much of the theoretical benefit provided by the underlying AI models.

Lack of Personalization and Custom Object Recognition Current state-of-the-art object detection frameworks, such as YOLO or standard SSD models, are typically trained on massive public datasets like COCO or ImageNet [17]. Consequently, they excel at identifying broad, generic categories—such as "chair," "car," or "person"—but fail to recognize specific items that are uniquely important to an individual user [18]. For example, a standard system cannot differentiate between "a water bottle" and "the user's specific medication bottle." Existing systems lack an intuitive mechanism for users to seamlessly register and retrieve their own personal belongings, fundamentally limiting their utility in managing personal spaces.

Fragmented Application Ecosystems Finally, the current market for assistive technology is highly fragmented [19]. Users are frequently forced to constantly switch between separate, isolated applications: one app for GPS navigation, another for reading text, and yet another for object recognition. This lack of integration forces visually impaired individuals to manage multiple audio streams and varied interface designs simultaneously, significantly increasing their cognitive load and causing digital fatigue [20]. Existing ecosystems rarely fuse these multimodal capabilities into a single, cohesive, context- aware platform.

IV. RESEARCH GAP

A critical analysis of the current literature reveals a substantial gap in the development of truly unified, multimodal assistive platforms for the visually impaired [7]. While extensive research has been devoted to optimizing isolated deep learning tasks— such as generic bounding-box object detection or voice-to-text translation—few studies address the holistic integration of these capabilities into a single, cohesive user experience [12]. Furthermore, there is a pronounced lack of focus on zero-click, secure accessibility, as most "hands-free" applications still inherently require physical interaction during critical setup or authentication phases [15]. Another significant gap exists in the domain of

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, March 2026
Impact Factor: 8.2

personalized object recognition, where existing models struggle to dynamically learn and identify specific, user-defined personal items without requiring extensive model retraining [17]. Current methodologies also frequently ignore the computational constraints of continuous video processing on mobile edge devices, leading to theoretical models that fail to maintain real-time performance or drain battery life in real-world scenarios [19]. Bridging these gaps requires the design of highly integrated frameworks, like SmartVision, which prioritize seamless biometric entry, low-latency contextual awareness, and comprehensive safety protocols over isolated technological demonstrations [21].

V. CONCLUSION

In conclusion, the SmartVision system demonstrates a significant advancement in the design and deployment of multimodal assistive technologies for the visually impaired .By unifying real-time scene description, generic object detection, and GPS-assisted navigation into a single, cohesive framework, the platform effectively reduces the cognitive burden typically associated with managing multiple isolated applications . The introduction of a completely hands-free, secure biometric authentication pipeline, alongside an optimized, on-demand personal object matching system, directly addresses critical accessibility barriers that have historically hindered user autonomy. Furthermore, the integration of a proactive, voice-activated safety infrastructure ensures that users remain protected in unpredictable environments [20]. Experimental deployment confirms that by intelligently balancing computational workloads between client-side capture and cloud-edge inference, the system maintains the low latency essential for real-world functionality without compromising accuracy . Ultimately, SmartVision not only bridges the existing gap between complex deep learning models and practical usability, but it also establishes a robust, holistic foundation for the future development of truly independent and human- centric assistive living platforms

VI. SUMMARY

From the analysis of prior research, we observe that assistive computer vision technologies have evolved significantly from traditional obstacle-detection mechanisms to advanced deep learning architectures capable of complex semantic understanding. Early techniques relying on simple sensors or handcrafted feature-based algorithms provided foundational progress for spatial awareness but lacked the robustness to accurately interpret dynamic, unconstrained real-world environments [2][3]. The emergence of deep Convolutional Neural Networks (CNNs) marked a major breakthrough by enabling automatic hierarchical feature learning, resulting in highly accurate object detection and scene classification capabilities [4][9]. However, these sophisticated deep learning models introduced immense computational costs and high latency, restricting their standalone deployment on the mobile and embedded platforms typically carried by visually impaired users [10][11].

Recent developments in lightweight architectures, such as MobileNet and compressed detection frameworks, attempt to balance speed and accuracy, enabling real- time edge processing while maintaining acceptable environmental awareness [13][16]. Transformer-based models further enhance global contextual understanding but require extensive training data and hardware resources, severely limiting their practical usability for instantaneous assistive feedback [12][18]. Additionally, while multimodal intelligent systems demonstrate the critical importance of combining visual data with continuous contextual signals— such as GPS and audio—many existing frameworks depend heavily on fragmented applications and isolated cloud infrastructures, raising significant latency and cognitive load concerns [15][17]. Overall, current research demonstrates strong progress in isolated detection and recognition capabilities but reveals persistent, systemic gaps in seamless biometric accessibility, edge efficiency, personalized object retrieval, and holistic, zero-click user interfaces [19][21]. These findings directly motivate the development of SmartVision as a unified, lightweight, and real-time framework that integrates accurate multimodal perception with accessible, highly deployable intelligence.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, March 2026
Impact Factor: 8.2

VII. ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our guide, Prof. Rashmi Mahajan (Project Guide), for her invaluable guidance, continuous support, and constructive suggestions throughout the course of this research work. Her insights and encouragement played a crucial role in the successful completion of this project. We are deeply thankful to Dr. Renuka Deshpande (Head of Department, AIML), Shivajirao S. Jondhale College of Engineering (SSJCOE), for providing the academic support, resources, and motivation necessary for carrying out this work effectively. We also extend our heartfelt appreciation to Shivajirao S. Jondhale College of Engineering (SSJCOE), Dombivli, for offering a conducive learning environment and the required facilities that enabled us to conduct this research successfully.

REFERENCES

[1] S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1195-1215, 2020.
[2] A. Tapu, B. Mocanu, and T. Zaharia, "Wearable assistive devices for the visually impaired: A state of the art survey," Computer Vision and Image Understanding, vol. 192, p. 102894, 2020.
[3] R. Manduchi and J. Coughlan, "(Computer) Vision without sight," Communications of the ACM, vol. 55, no. 1, pp. 96-104, 2012.
[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.
[5] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in International conference on neural information processing, Springer, Berlin, Heidelberg, pp. 117-124, 2013. (Reference for FER-2013 dataset)
[6] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200-205, 1998. (Reference for JAFFE dataset)
[7] A. Alqahtani, A. Al-Tariq, and S. M. Qaisar, "Real-time object detection and recognition for the visually impaired using edge computing," IEEE Access, vol. 9, pp. 125197-125208, 2021.
[8] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1701-1708, 2014.
[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788, 2016.
[10] S. S. A. Zaidi et al., "A survey of modern deep learning based object detection models," Digital Signal Processing, vol. 126, p. 103514, 2022.
[11] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.
[12] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
[13] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510-4520, 2018.
[14] P. P. Jain et al., "Smart interactive cane for the visually impaired utilizing IoT and cloud architecture," IEEE Internet of Things Journal, vol. 7, no. 10, pp. 9504-9515, 2020.
[15] T. Baltrušaitis, C. Ahuja, and L. P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423-443, 2018.
[16] E. Li, L. Zeng, Z. Zhou, and X. Chen, "Edge AI: On-demand accelerating deep neural network inference via edge computing," IEEE Transactions on Wireless Communications, vol. 19, no. 1,
pp. 447-457, 2019.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, March 2026
Impact Factor: 8.2

[17] Y. Bai, Y. Zhang, Y. Wang, and F. Li, "Context-aware navigation for visually impaired people in dynamic environments," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 27, no. 8, pp. 1629-1638, 2019.
[18] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," arXiv preprint arXiv:1510.00149, 2015.
[19] A. K. Jain, K. Nandakumar, and A. Ross, "50 years of biometric research: Accomplishments, challenges, and opportunities," Pattern Recognition Letters, vol. 79, pp. 80-105, 2016.
[20] H. Rashed, S. El-Din, A. Elsayed, and R. E. A. El-Sehiemy, "Secure cloud-based framework for emergency tracking and distress signaling in IoT networks," IEEE Access, vol. 8, pp. 110467-110481, 2020.
[21] X. Wang, Y. Zhang, L. Nie, and Q. Tian, "Real-time multimodal interface for assistive technologies: Balancing latency and cognitive load," IEEE Transactions on Multimedia, vol. 23, pp. 3201-3212, 2021