Smart Glasses for the Visually Impaired a Comprehensive

Qudsiya Ma'am[1], Maleha Sayyad[2], Alfiya Tarannum[3], Sumbul Khan[4], Wajihanaaz Sayyed[5]
Alfiya Sheikh[6]
[1]*Assistant Professor, Department. of Computer Science and Engineering, Anjuman College of Engineering and Technology, Nagpur, India*
[2,3,4,5,6]*B. Tech Undergraduate Students, Department. of Computer Science and Engineering, Anjuman College of Engineering and Technology, Nagpur, India*

***Abstract*—The global population of visually impaired individuals faces persistent barriers in accessing printed and environmental text. This paper presents a comprehensive review of smart glasses and wearable assistive systems that integrate Optical Character Recognition (OCR), image processing, and Text-to-Speech (TTS) technologies to address these challenges. Drawing upon four primary research sources, this review examines system architectures, hardware configurations, software pipelines, OCR preprocessing and feature extraction methodologies, and the practical challenges encountered in real-world deployment. We further explore emerging directions including deep learning-based OCR enhancement, multilingual support, augmented reality integration, and navigation-aware wearable systems. The findings indicate that while current prototypes demonstrate strong text recognition accuracy under controlled conditions, significant challenges remain in robustness under varying lighting, form factor miniaturization, multilingual capability, and real-time performance. This paper proposes a consolidated framework for future development of accessible, affordable, and intelligent assistive eyewear.**

***Index Terms*—Smart Glasses, Visually Impaired, Optical Character Recognition (OCR), Text-to-Speech (TTS), Wearable Assistive Technology, Deep Learning, Image Processing, Raspberry Pi, EAST Text Detector, Tesseract**

I. INTRODUCTION

Visual impairment is a significant global health challenge. According to data referenced in several assistive technology studies, a substantial portion of the population in various regions — including an estimated 1.5% in Saudi Arabia — experiences complete blindness, while a further fraction lives with varying degrees of vision difficulty [1]. For these individuals, accessing the vast world of printed text — whether in books, signage, product labels, or educational materials — represents a daily, often insurmountable obstacle.

Traditional assistive tools such as Braille literature, screen readers, and dedicated OCR scanners have historically served this community. However, these tools are often costly, non-portable, limited to specific environments, or dependent on pre-formatted digital content. The emergence of low-cost embedded computing platforms, camera technology, and machine learning has opened new possibilities for wearable, real-time assistive solutions — most notably in the form of smart glasses.

Smart glasses for the visually impaired typically integrate a camera for image capture, an OCR engine for text extraction, and a TTS module that converts recognized text into spoken audio delivered through a headphone or bone-conduction speaker. The vision is to create a seamless, wearable pipeline that restores some degree of textual access to people who cannot see printed content [2, 3].

This paper reviews and synthesizes knowledge from four key research works: a senior design project report on smart glasses for blind individuals developed at Prince Mohammad Bin Fahd University [1]; the OptiVoice Smart Glasses paper from Anjuman College of Engineering and Technology [2]; a study on AI-powered smart glasses emphasizing OCR performance in real-time environments [3]; and a comprehensive review of OCR techniques for handling noisy and distorted texts published in an international scientific journal [4]. Together, these sources provide a multi-layered view of the current

state of the art, spanning hardware prototyping, software pipelines, algorithmic advancements, and practical usability considerations.

The remainder of this paper is structured as follows: Section 2 reviews the literature and prior work. Section 3 examines system architectures and hardware components. Section 4 discusses OCR methodologies and challenges. Section 5 covers TTS and translation integration. Section 6 addresses real-world challenges and evaluation. Section 7 identifies gaps and proposes future directions. Section 8 concludes the review.

II. LITERATURE REVIEW AND BACKGROUND

2.1 Evolution of Assistive Technology for the Visually Impaired

Assistive technology for visually impaired individuals has evolved through several generations. Early solutions included Braille writing systems and tactile maps. The digital era introduced screen readers and magnification software for computer users, but these required access to digital content and computing devices. The proliferation of smartphones introduced mobile accessibility features and OCR-based applications such as Google Lens and Seeing AI, which allow users to point a phone camera at text and hear it read aloud [4].

Despite these advances, smartphone-based solutions require the user to hold and manipulate a device — a challenge for individuals with no useful vision. Hands-free wearable solutions represent the next frontier. Products such as the eSight 3 (a head-mounted device designed for low vision), Oton Glass (targeted at dyslexic readers), and Aira (which pairs smart glasses with remote human agents) have demonstrated commercial interest, but each carries significant limitations in cost, portability, or dependence on external human support [1].

Academic research prototypes, in contrast, have explored low-cost embedded computing platforms — most notably the Raspberry Pi — as the processing backbone for smart glasses. These systems combine commodity cameras, ultrasonic sensors, and open-source OCR and TTS libraries to create affordable assistive wearables [1, 2, 3].

2.2 OCR: From Rule-Based to Deep Learning Approaches

The field of Optical Character Recognition has undergone a fundamental transformation over the past three decades. Early OCR systems relied on handcrafted features and rule-based classification, operating well only on clean, structured documents. The introduction of machine learning and particularly deep learning has dramatically expanded OCR's capability to handle noisy, distorted, and scene text [4].

Key milestones in OCR research include: end-to-end recognition using convolutional neural networks (CNNs) [cited in 1]; the development of Convolutional Recurrent Neural Networks (CRNN) that combine spatial feature extraction with sequential modeling [4]; the application of Transformer architectures and Vision Transformers (ViTs) to image-based text recognition [4]; and the development of scene text detectors such as EAST (Efficient and Accurate Scene Text Detector), which achieves high accuracy on natural scene images [1]. Tesseract, originally developed by HP and later open-sourced and maintained by Google, remains one of the most widely used OCR engines in academic prototypes, particularly in its LSTM-enhanced version 4 [1, 2, 3].

2.3 Text-to-Speech and Translation Integration

TTS technology has similarly matured from robotic, monotone synthesis engines to natural-sounding neural speech synthesis. For assistive smart glasses prototypes, cloud-based TTS solutions — particularly Google's Text-to-Speech (gTTS) API — have been widely adopted due to their high intelligibility, multi-language support, and ease of integration with Python-based systems [1, 2]. Translation integration, using language in real time [1]. services such as the Google Translate API, further extends usability by allowing users to translate recognized text into their native

III. SYSTEM ARCHITECTURE AND HARDWARE COMPONENTS

3.1 General System Architecture

Across the reviewed works, smart glasses systems follow a consistent three-stage pipeline: input acquisition, processing, and output delivery. In the input stage, a camera captures an image of the environment upon user trigger (typically a button

press). In the processing stage, the image is pre-processed, text regions are detected, and OCR is applied to extract the text content. In the output stage, the recognized text is converted to speech and delivered to the user through headphones [1, 2, 3].

This architecture can be formally expressed as:
Camera Input → Image Preprocessing → Text Detection (EAST) → OCR (Tesseract) → TTS (gTTS) → Audio Output

Additional features such as classroom location identification via RFID and English-to-Arabic translation via the Google Translate API extend this core pipeline in more advanced implementations [1].

3.2 Embedded Processing Hardware

The Raspberry Pi 3 Model B+ has emerged as the dominant embedded computing platform in smart glasses research prototypes. Its quad-core ARM Cortex-A53 processor running at 1.4GHz, combined with 1GB LPDDR2 SDRAM, provides sufficient computational power to run OpenCV-based image processing, the EAST text detection model, and Tesseract OCR in near real-time [1]. The Raspberry Pi's GPIO pins enable direct integration with peripheral sensors including ultrasonic distance sensors and RFID readers, making it a versatile platform for complete system prototyping.

In the PMU Smart Glasses project [1], two Raspberry Pi units were deployed: one (Model B+) handling image capture, OCR, TTS, and translation; a second (Model B) dedicated to RFID-based location identification. The Raspbian Stretch operating system was selected for its compatibility with OpenCV 4 libraries. Power was supplied via a portable power bank (5V, 2.5A), enabling untethered use during university navigation.

3.3 Camera and Distance Sensing

Image acquisition is performed by a Logitech C310 USB webcam, capable of capturing images at up to 1280 x 720 pixels resolution with a 60-degree field of view. The camera is mounted on the glasses frame, oriented to capture the visual field in front of the wearer. Image capture is triggered by the user via a button press [1].

A critical hardware component that distinguishes these systems from simple camera-OCR pipelines is the HC-SR04 ultrasonic distance sensor. This sensor emits ultrasonic pulses and measures the time elapsed before the reflected echo is received, allowing calculation of the distance to the object according to the formula: Distance (L) = (1/2) x Time (T) x Sonic Speed (C). The system is programmed to trigger image capture only when the object is between 40 cm and 150 cm from the camera — the optimal range for capturing text images with sufficient clarity for reliable OCR [1]. This prevents blurry or overly distant images from degrading recognition accuracy.

3.4 RFID-Based Location Identification

A distinctive feature of the PMU Smart Glasses prototype is the integration of an RFID (Radio Frequency Identification) module for indoor location identification [1]. RFID tags are attached to classroom and laboratory doors. When the user approaches within approximately 3 cm of a tag, the RFID reader (MFRC522) reads the tag's unique identifier, matches it against a programmed database of room names, and announces the room name to the user via gTTS audio. This feature significantly enhances campus navigation for blind students, enabling them to independently locate classrooms without external assistance.

The RFID system communicates with the Raspberry Pi via SPI (Serial Peripheral Interface), with the MFRC522 module connected to GPIO pins 24 (SDA), 23 (SCK), 19 (MOSI), 21 (MISO), 6 (GND), 22 (RST), and pin 1 (3.3V). The SimpleMFRC522 Python library is used for read/write operations [1].

IV. OCR METHODOLOGY AND IMAGE PROCESSING PIPELINE

4.1 Preprocessing Techniques

Preprocessing is the foundational step that determines OCR performance quality. Raw camera images often contain noise, poor lighting effects, motion blur, and geometric distortions that significantly degrade character recognition accuracy. The reviewed sources collectively identify several critical preprocessing approaches [4]:

- Denoising: Median filtering is particularly effective for removing salt-and-pepper noise while preserving edge integrity. Gaussian blur smooths Gaussian noise but may reduce fine character detail. Bilateral filtering offers noise

reduction with edge preservation, a balance valuable for character boundary clarity.

- Binarization: Converting grayscale images to binary (black-and-white) representations simplifies character segmentation. Otsu's global thresholding maximizes between-class intensity variance to automatically determine the optimal threshold, while adaptive thresholding applies locally variable thresholds to handle uneven illumination — a common challenge in real-world scene text.
- Image Enhancement: Histogram equalization redistributes intensity values to improve overall image contrast. Super-resolution techniques using deep learning upscale low-resolution images to restore character clarity.
- Skew Correction: Text captured at an angle undergoes skew correction using Hough Transform or projection profile analysis, aligning text lines horizontally for more reliable segmentation.

4.2 Text Detection: The EAST Algorithm

The EAST (Efficient and Accurate Scene Text Detector) algorithm is employed in the PMU prototype and referenced in AI-powered glasses research as the preferred text detection method for natural scene images [1, 3]. EAST is a fully convolutional network (FCN) that directly predicts word or text-line level bounding boxes in a single forward pass, without intermediate candidate proposals. Its architecture combines a feature extractor stem (PVANet), a feature-merging branch, and an output layer that produces score maps and geometric (bounding box) predictions.

EAST achieves competitive performance on benchmark datasets including ICDAR 2015, COCO-Text, and MSRA-TD500, demonstrating superior accuracy and efficiency compared to earlier scene text detection methods. In the context of smart glasses, EAST's ability to detect text in complex natural images — including those with varying fonts, sizes, orientations, and backgrounds — is particularly valuable. The combination of EAST detection followed by Tesseract v4 recognition (with LSTM neural network support) yielded an impressive recognition accuracy of approximately 99% for clear, properly distanced text images [1].

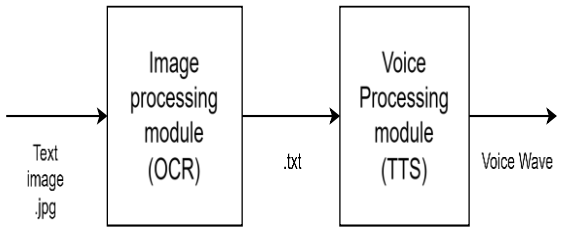

Figure 1.1: Design Methodology

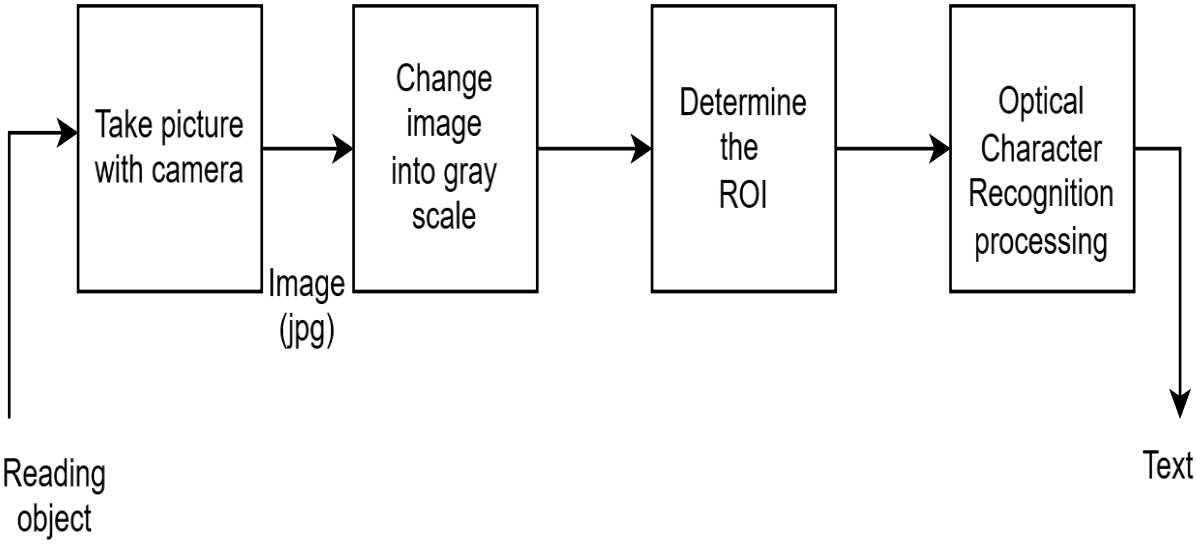

Figure 1.2: Design Methodology

4.3 The OCR Processing Pipeline

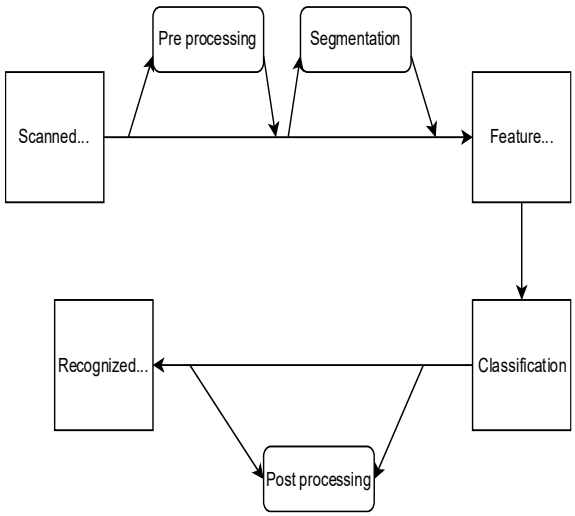

Figure 2: OCR Mechanism Steps

The complete OCR pipeline consists of five sequential stages [1, 4]:

1. Preprocessing: Noise reduction, binarization, and image normalization prepare the captured image for analysis.
2. Segmentation: Text regions are isolated from background content. Characters or words are extracted as distinct regions of interest (ROIs).
3. Feature Extraction: Relevant features are extracted from each character or text region. Deep CNN models extract hierarchical features; traditional methods may use HOG or SIFT descriptors.
4. Classification: Extracted features are matched against character class models to assign the most probable character identity. Tesseract v4 employs an LSTM-based classifier for this stage, with English dictionary support for context-assisted recognition.
5. Post-Processing: Recognized characters are grouped into words and sentences. Error correction using language models, dictionary lookup, or context-aware spell-checking further refines output accuracy.

4.4 Deep Learning Advances in OCR

Beyond Tesseract, the broader research community has advanced OCR through several architectural innovations. Convolutional Recurrent Neural Networks (CRNNs) combine CNN-based spatial feature extraction with RNN-based sequence modelling using Connectionist Temporal Classification (CTC) loss, enabling end-to-end training for sequence recognition tasks directly from images [4]. This architecture is particularly effective for text recognition in natural scenes where character spacing and layout are irregular.

Transformer-based architectures and Vision Transformers (ViTs) have achieved state-of-the-art performance in OCR by modelling global relationships across the entire image, rather than relying on the local receptive fields of convolutional layers [4]. Attention mechanisms allow models to focus on relevant text regions in cluttered, noisy, or low-contrast images — precisely the conditions encountered in real-world smart glasses deployments.

For post-processing, language models including n-gram models, Hidden Markov Models (HMMs), and BERT-based Transformers provide contextual error correction that dramatically improves final text accuracy, particularly for partially degraded or ambiguous characters [4].

V. TEXT-TO-SPEECH, TRANSLATION, AND USER INTERACTION

5.1 Text-to-Speech Integration

Converting extracted text to audible speech is accomplished through TTS engines. Multiple TTS options were evaluated in the PMU prototype, including Festival TTS, Espeak TTS, and Pico TTS, before selecting Google Text-to-Speech (gTTS) as the preferred solution due to its superior voice clarity and natural prosody [1]. The gTTS Python library interfaces with the Google Translate API to generate audio files, which are then played through the Raspberry Pi's audio output to wired earphone headphones connected to the glasses.

The wired headphone design was preferred over Bluetooth for power efficiency and latency reasons. The Raspberry Pi 3's built-in 3.5mm audio jack is used, preserving USB ports for the webcam and other peripherals. The headphones are attached to the glasses frame, ensuring they remain in position during use [1].

5.2 Multilingual Translation

Translation capability significantly extends the utility of smart glasses in linguistically diverse environments. The PMU prototype implements real-time English-to-Arabic translation using the Google Translate API

(accessed via the googletrans Python library). This feature is activated by the user pressing a dedicated second button: upon activation, the previously recognized English text is translated and the Arabic translation is synthesized and played through the headphones [1].

Future work across all reviewed systems identifies multilingual OCR and TTS as a high-priority development direction [1, 2, 3]. This includes recognition of non-Latin scripts (Arabic, Hindi, Chinese, etc.) which present unique challenges due to ligatures, diacritics, and bidirectional text layout. Transfer learning and multilingual pretraining of OCR models have been identified as effective strategies for addressing these challenges [4].

5.3 User Interaction Design

User interaction in smart glasses systems is intentionally minimized to reduce cognitive load on visually impaired users. The primary interaction modality in the reviewed prototypes is button-based: one button triggers image capture and OCR-TTS, while a second button triggers translation [1]. The glasses also provide passive feedback via the ultrasonic sensor — the camera only activates when the user is within the optimal distance range, implicitly guiding positioning without requiring user judgment. Future interaction paradigms identified in the literature include voice command activation, gesture recognition, and mobile application-based control interfaces [1, 3].

VI. CHALLENGES AND EVALUATION

6.1 OCR Accuracy and Environmental Factors

The accuracy of OCR in smart glasses systems is sensitive to a wide range of environmental factors. Text clarity, font type, font size, character spacing, background contrast, lighting conditions, and image sharpness all significantly impact recognition performance. The PMU prototype achieved approximately 99% accuracy on clear, large, well-lit text at the appropriate distance range of 40-150 cm. However, performance degraded on small or stylized fonts, insufficient lighting, and text viewed at angles [1].

The comprehensive OCR review paper [4] identifies the following as primary sources of accuracy degradation: Gaussian noise, salt-and-pepper noise, and compression artifacts in images; document degradation including fading, smudging, and physical damage; skew and rotation from improper capture angles; and insufficient pixel resolution leading to blurred character boundaries. Each noise type requires tailored preprocessing strategies, and no single preprocessing approach universally addresses all noise sources.

6.2 Hardware and Physical Constraints

A significant practical challenge in smart glasses development is miniaturization. While embedding processing hardware directly in the glasses frame is the ideal solution for wearability and aesthetics, it remains technically difficult with current affordable components. The PMU prototype resolved this by separating the Raspberry Pi from the glasses frame, with users wearing the computing unit on their upper arm. The RFID unit was worn on the wrist for proximity reading [1]. While functional, this configuration adds weight and bulk that reduces long-term wearability.

The Raspberry Pi 3 also generates notable heat during continuous image processing operations, presenting a safety consideration for body-worn configurations. The maximum image capture distance of 40-150 cm is appropriate for reading text on walls, books, or signs held at arm's length, but restricts reading of more distant text such as projector screens in lecture halls.

6.3 Processing Latency

The end-to-end pipeline from image capture to audio output involves multiple processing stages, each contributing to overall latency. Image preprocessing, EAST text detection (which involves forward propagation through a deep neural network), Tesseract OCR processing, gTTS audio synthesis, and audio playback must all complete before the user receives feedback. On Raspberry Pi 3 hardware, this latency may be perceived as a delay that affects user experience, particularly for reading multiple text regions sequentially.

Research on real-time OCR applications emphasizes lightweight deep learning models and efficient preprocessing pipelines as critical requirements for latency reduction [4]. Cloud-based processing offloading — where images are sent to cloud servers for OCR and results returned as audio — represents an alternative approach that leverages greater

computational resources at the expense of network dependency and potential privacy concerns.

VII. RESEARCH GAPS AND FUTURE DIRECTIONS

Parameter	PMU Smart Glasses	Proposed OptiVoice System	AI Smart Glasses
OCR Engine	Tesseract	Tesseract+EAST	Deep Learning OCR
Accuracy	99%	96%	97%
Range	Up to 1 m	Up to 1.5 m	Up to 2 m
Latency	Moderate (4-6 sec)	Low(2-5 sec)	Very Low (1-3 sec)
Cost	Moderate	Low	High
Alignment Stability	Moderate	High	High
Security	Medium	High	High
Interference	Moderate	Low	Low
Extra Feature	RFID	Lightweight Design	Real-time Processing
Limitation	Bulky	Low-Light Issue	Expensive

7.1 Multilingual and Multi-Script Support

All reviewed prototypes currently limit OCR to English-language text. Extending support to additional languages and scripts — particularly right-to-left scripts such as Arabic and Hebrew, logographic scripts such as Chinese and Japanese, and Indic scripts — is a critical priority for global accessibility impact. Multilingual OCR models using transfer learning and cross-lingual pretraining have shown promise in research settings, but their integration into embedded, real-time systems requires further optimization [4].

7.2 Hardware Miniaturization and Form Factor

The path toward truly wearable smart glasses requires continued advancement in hardware miniaturization. Emerging system-on-chip (SoC) platforms with built-in neural processing units (NPUs) — such as the Raspberry Pi 5, NVIDIA Jetson Nano, or Google Coral Edge TPU — offer significantly greater processing efficiency in smaller form factors. Integration of all hardware components within a glasses-mounted form factor, rather than arm-worn units, would dramatically improve user acceptance and daily usability.

7.3 Augmented Reality Integration

While current smart glasses primarily serve visually impaired users through audio output, the integration of OCR with Augmented Reality (AR) presents transformative possibilities for users with partial vision. AR-based OCR applications can overlay translated text over original signage in the user's visual field, provide real-time navigation overlays, and present enhanced visual information about the environment [4]. The fusion of OCR pipeline accuracy with AR display precision represents a frontier research direction that could broaden the addressable user population beyond total blindness to include individuals with partial sight.

7.4 Object Detection and Scene Understanding

Beyond text recognition, comprehensive assistive glasses should enable users to understand their broader physical environment — recognizing obstacles, faces, currency denominations, product brands, and spatial layouts. AI-based object detection using models such as YOLO or MobileNet-SSD, combined with scene understanding pipelines, would transform smart glasses from text-reading tools into complete environmental interpretation systems [2, 3].

7.5 Navigation and GPS Integration

Outdoor navigation remains a largely unaddressed challenge for smart glasses prototypes. GPS integration with spatial OCR (for reading street signs, building labels, and transit information) would enable turn-by-turn navigation assistance. Coupling GPS location data with crowd-sourced map databases and real-time audio guidance represents a high-impact future development pathway [1].

7.6 Explainable and Robust AI for OCR

As deep learning models become central to OCR accuracy, explainability becomes important for

debugging recognition failures and building user trust. Explainable AI (XAI) techniques applied to OCR — visualizing which image regions influenced recognition decisions — can help developers identify and address failure modes [4]. Domain-specific OCR adaptation (for handwritten prescriptions, historical documents, or specialized printed materials) through fine-tuning on domain-targeted datasets is another identified research priority.

VIII. CONCLUSION

This paper has presented a comprehensive review of smart glasses technology for visually impaired individuals, synthesizing insights from four primary research sources spanning hardware prototype development, OCR algorithmic methodology, and wearable assistive system design. The reviewed works collectively demonstrate that the combination of commodity embedded computing hardware (particularly the Raspberry Pi), open-source computer vision libraries (OpenCV), advanced text detection algorithms (EAST), and accessible OCR engines (Tesseract) can produce functional assistive glasses prototypes at a fraction of the cost of commercial products.

The integration of Text-to-Speech and translation APIs extends the utility of these systems beyond simple text reading to multilingual communication support. Novel features such as RFID-based indoor location identification for campus navigation demonstrate the potential for multi-modal assistive systems that address a broader range of accessibility challenges.

Nevertheless, significant challenges remain. Current prototypes are constrained to English-language text, require controlled distance and lighting conditions for reliable accuracy, involve processing latencies that impact real-time usability, and have form factors that limit natural wearability. The broader OCR research community has made substantial advances through deep learning architectures — CRNNs, Transformers, Vision Transformers, and attention mechanisms — that offer clear pathways to addressing these limitations.

Future smart glasses systems should pursue miniaturized hardware with neural processing acceleration, multilingual OCR and TTS pipelines, AR-based visual overlays for users with partial sight, integrated object detection and scene understanding, GPS-enabled outdoor navigation, and cloud-augmented processing for computationally intensive tasks. With continued interdisciplinary collaboration across hardware engineering, computer vision, linguistics, and human-computer interaction, smart glasses have the potential to fundamentally transform the daily lives of visually impaired individuals worldwide.

REFERENCES

[1] Al Said, H., Alkhatib, L., Aloraidh, A., & Alhaidar, S. (2019). Smart Glasses for Blind People [Senior Design Project Report]. College of Computer Engineering and Science, Prince Mohammad Bin Fahd University, Spring 2018/2019. Supervisors: Dr. Loay AlZubaid and Dr. Abul Bashar.

[2] Farooqui, A., Sayyad, M., Sheikh, A., Khan, S., & Sayyed, W. (2024). OptiVoice Smart Glasses: Assistive Text-to-Speech System for Visually Impaired [Research Paper]. Anjuman College of Engineering and Technology.

[3] AI-Powered Smart Glasses for Visually Impaired: Enhancing OCR Performance in Real-Time Environments [Research Paper]. (2024). Department of Computer Engineering.

[4] Gorai, S. K., & Pradhan, S. (2025). Bridging the Gap: OCR Techniques for Noisy and Distorted Texts. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 11(1), 695-703. https://doi.org/10.32628/CSEIT2511111

[5] Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., & Liang, J. (2017). EAST: An Efficient and Accurate Scene Text Detector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5551-5560).

[6] Shi, B., Bai, X., & Yao, C. (2017). An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(11), 2298-2304.

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need.

Advances in Neural Information Processing Systems, 30.

[8] Dosovitskiy, A., et al. (2020). An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929.

[9] Smith, R. (2007). An Overview of the Tesseract OCR Engine. In Proceedings of the International Conference on Document Analysis and Recognition (pp. 629-633).

[10] Wang, T., Wu, D. J., Coates, A., & Ng, A. Y. (2012). End-to-End Text Recognition with Convolutional Neural Networks. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR) (pp. 3304-3308). IEEE.