IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141298

A Real-Time Multimodal Assistive Framework Integrating Ensemble OCR, Object Detection, Text Analytics, and Haptic Feedback

Dr. T. R. Muhibur Rahman[1], Prashanth J[2], Karnatakam Sai Anirudh[3], Jagat Singh[4], Haseeb Ahmed S[5]

Dept. Of CSE, Ballari Institute of Technology and Management, Ballari, India[1]

7th Semester B.E. (CSE), Ballari Institute of Technology and Management, Ballari, India[2-5]

Abstract: Smart Reader is designed as an accessibility-focused system that supports non-visual reading by combining text recognition, scene understanding, audio guidance, and tactile cues. The platform utilizes a pair of complementary OCR engines in concert with a lightweight object-analysis module to interpret both text and surrounding context from incoming video frames. Such an optimized backend, implemented on top of Fast API, empowers the pipeline to handle each frame within about two hundred milliseconds, which can enable several frames per second on regular CPU hardware. A specially designed mapping function transforms the OCR characteristics, including confidence levels, text density, and semantic weight, into structured vibration signals that help users haptically perceive document layout. Experimental studies have shown higher recognition reliability than using a single OCR method alone, along with consistent detection quality and responsive system behavior. In general, Smart Reader provides an improved pathway for visually impaired users to access printed or on-screen information based on a powerful combination of perception, interpretation, and haptic assistance.

Keywords: OCR, Haptic Feedback, YOLOv8, Assistive Technology, Fast API, Real-Time Processing.

I. INTRODUCTION

For people with visual impairments, accessing text that is printed or on-screen continues to be a major barrier. The existing tools predominantly rely on spoken output; unfortunately, audio alone often cannot convey how information is structured on a page. Structural cues such as headings, tables, or the spatial relationships between sections become linearized into a flat stream of speech, making it hard to perceive hierarchy or find critical parts of a document. Not having a secondary feedback channel leaves users without a clear sense of how the content is organized, slowing down navigation and reducing independence.

Although recent improvements in vision models and lightweight inference have increased the capability of assistive applications, everyday usage still reveals several deficiencies. Systems designed to read text from a camera feed tend to break down when exposed to uncontrolled lighting, sharp or tilted capture angles, cluttered scenes, or mixed fonts and sizes. The same issues translate into inconsistent recognition and unpredictable reading output. Typical speech engines, however, treat every word with exactly the same prosody, in that emphasis, structural markers, and contextual importance are rarely communicated.

While widely employed in other interaction domains, touch-based feedback has not yet been fully explored as a complement to the translation of visual into audio. Vibrations and tactile patterns can convey information related to transitions, grouping, or relative importance without interfering with speech and remain effective when it is not easy to hear audio or when audio needs to be discreet. Well-designed haptic cues can serve as a fast parallel channel in helping users form a mental model of the document, enhancing comprehension by means of immediate, non-verbal signaling.

II. RELATED WORK

R. Smith [1] gave a description of the internal design of the Tesseract OCR engine in detail and discussed its move into an open-source recognition system. This work describes the connected-component analysis, character segmentation, and adaptive classification that Tesseract applies to printed text recognition, including complications brought in by font variations, spacing, and document noise. This seminal work placed Tesseract as one of the most

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025
DOI: 10.17148/IJARCCE.2025.141298

reliable baseline OCR engines at that time and appears to continue influencing modern text recognition systems.

Baek et al. [2] discuss inconsistencies in the results of scene text recognition models across different studies. They suggested a standardized evaluation pipeline that minimizes variances due to dataset selection, training settings, and preprocessing. Benchmarking several architectures on identical experimental settings, this paper exposed misleading performance claims in earlier studies and thus highlighted the need for unified testing. This will go a long way in developing stable and comparable OCR models, one of them being EasyOCR

A. Bochkovskiy et al. [3] introduce the object detection model YOLOv4, which, with new architectural components combined with training strategies, is optimized to further improve accuracy without losing speed. Techniques such as the use of CSPDarknet53, weighted residual connections, and optimized data augmentation make this work state-of-the-art. YOLOv4 performs very well on benchmark datasets and is suitable for resource- constrained environments; therefore, it will be relevant for real-time visual detection in assistive systems.

Khusro et al. [4] review the state of the art in haptic feedback technologies that assist visually impaired users in indoor navigation. The authors classify various types of tactile systems, vibration patterns, and wearable devices that communicate spatial information in this current paper. Then, they analyze the strengths and limitations of those approaches, pointing out how this could be used to complement auditory guidance through haptic cues. They conclude that with proper integration into assistive tools, haptic feedback has the potential to significantly enhance the environmental awareness and independence of a blind user.

Hutto et al. [5] propose VADER, a rule-based sentiment analysis system tuned for social media and short textual content. In this model, a handcrafted lexical dictionary, along with a set of heuristics, such as but not limited to intensity modifiers, negation handling, and punctuation rules, is used in order to create high-accuracy sentiment scores. It was intended to be light and interpretable, and strong performance is hence achieved without the need for large-scale training data, which makes it useful in real-time sentiment analysis tasks.

Rose et al. [6] present several methodologies applied in automatic keyword extraction and discuss how such methods enable other applications, including text summarization, topic identification, and information retrieval. Their work compares the statistical and linguistic approaches, together with machine learning-based approaches, using examples to provide practical applications. The authors underline the identification of key phrases as the first stage of any downstream analytics and consider keyword extraction as part of each modern text-processing system.

Hersh [7] provides an overview of many types of assistive technologies that have been developed to facilitate people with visual impairments regarding mobility, reading systems, and navigation aids, as well as access software. Further, it considers design issues, user needs, and practical implementation problems arising when those techniques are applied in practice. The book is a big contribution in the accessibility technologies area and remains a guiding force for research on solutions for visually impaired persons.

Shi et al. [8] propose a design that couples CNN feature extraction with recurrent sequence modeling for end-to-end scene text recognition via a CRNN. The network takes as input the whole line of text, considering it as a sequence without the necessity of character-level segmentation. This gives much better robustness to all kinds of distortions, irregular shapes, and diversified fonts. In this paper, a CTC loss function was utilized to align predictions over the input sequences, allowing efficient training for precise recognition in natural scene texts. Indeed, most of the modern OCR systems, including EasyOCR applied in our work, are based on this model.

Redmon et al. [9] introduce YOLOv3-a very efficient real-time object detection model using Darknet-53 for multi-scale predictions. The improvements of this paper result in a large increase in speed and accuracy compared to its previous versions, placing YOLO among the strongest solutions for fast visual detection tasks relevant in assistive systems.

Grootendorst [10] proposes KeyBERT: a simple keyword extraction method which leverages BERT embeddings to find phrases which are most semantically similar to a document. The generated keywords are more meaningful compared to statistical models. This technique also aligns with embedding-based keyword extraction used in our system.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141298

III. PROPOSED SYSTEM

Smart Reader is the proposed AI-enabled multimodal document understanding framework that realizes high- accuracy text extraction with real-time assistance to the visually impaired. The proposed system contains a multi-stage preprocessing pipeline, hybrid ensemble OCR model, object detection with YOLOv8, and advanced text analytics to produce accurate, fast, and context-aware results.

It accepts input from images, PDFs, or live camera frames, does some preprocessing by enhancing through grayscale conversion and adaptive thresholding, after which the resulting images are sent for processing through an ensemble of Tesseract and EasyOCR, picking the best based on confidence scoring and semantic validation. If scene understanding is enabled, it detects objects around it using YOLOv8 and generates natural-language descriptions.

This output is further analyzed for keywords, sentiment analysis, readability scoring, and optionally semantic analysis with Gemini. The final output is provided as text or converted to speech with an integrated TTS engine; therefore, it is accessible for blind users. It also provides support for real-time scanning with haptic feedback to enable the user to properly set the camera.

It couples multi-engine OCR, AI analytics, and object detection in one single optimized pipeline; hence, the overall accuracy is higher with lesser latency and more accessibility.

IV. METHODOLOGY

The architecture of the SmartReader system relies on a module-based workflow that is designed to improve OCR accuracy, reduce processing latency, and further enhance the accessibility for visually impaired users by using a methodology that integrates multi-step preprocessing, adaptive ensemble OCR, semantic correction, object detection, and integrated text analytics.

A. Preprocessing Module

Preprocessing of the input images/PDF pages is carried out by converting them to grayscale, performing Otsu thresholding, and applying adaptive thresholding. This allows for multiple enhanced versions of an image to provide more diverse input to the OCR engines with the aim of achieving higher robustness.

B. Adaptive Ensemble OCR Module

SmartReader runs both Tesseract and EasyOCR in parallel, and each engine processes its own pre-processed image variant. After recognition, every engine's output is assigned an average confidence score:

$$C_i = \frac{1}{n} \sum_{k=1}^{n} \text{conf}_k \quad (1)$$

The system then selects the most confident OCR result:

$$\text{OCR}_{\text{final}} = \arg\max_i (C_i) \quad (2)$$

This is an ensemble approach that picks dynamically, for any given input, the most reliable output; it naturally outperforms any single OCR model in terms of overall character and word accuracy.

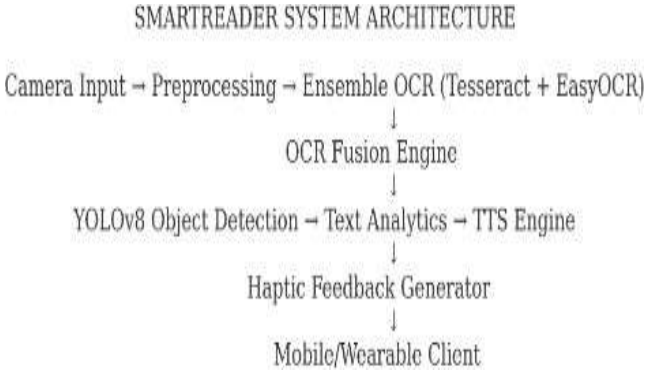

Figure 1. System-level illustration of the SmartReader pipeline, showing how input images move through preprocessing, OCR engines, detection modules, analytics, and the final output layer

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141298

C. Semantic Correction Module

First, the OCR text is refined by computing the sentence embeddings using SBERT. The similarity between the extracted text and its contextual reference is determined as:

$$\text{Cosine Similarity} = \frac{E_{\text{ocr}} \cdot E_{\text{ref}}}{|E_{\text{ocr}}| \; |E_{\text{re}}|} \tag{3}$$

Low similarity segments are corrected to increase coherence; the semantic errors induced by OCR are reduced.

D. Object Detection Module

YOLOv8 is used to detect scene objects at runtime. Detections with confidence above a threshold are retained:

$$\text{Conf}_{\text{obj}} = \sigma(f_{\text{cls}} + f_{\text{bbox}}) \tag{4}$$

This enables the system to provide environmental awareness to visually-impaired users alongside the reading of text.

E. Text Analytics Module

Extracted text is further analysed for keywords, readability, and context. TF-IDF is applied for keyword scoring:

$$\text{TFIDF}(t, d) = \text{TF}(t, d) \cdot \log \frac{N}{\text{df}(t)} \tag{5}$$

This improves information extraction and enhances the system's utility for document understanding.

F. Text-to-Speech Module

The final processed text is converted to audio using gTTS or Pyttsx3, having in mind further access when network conditions are poor.

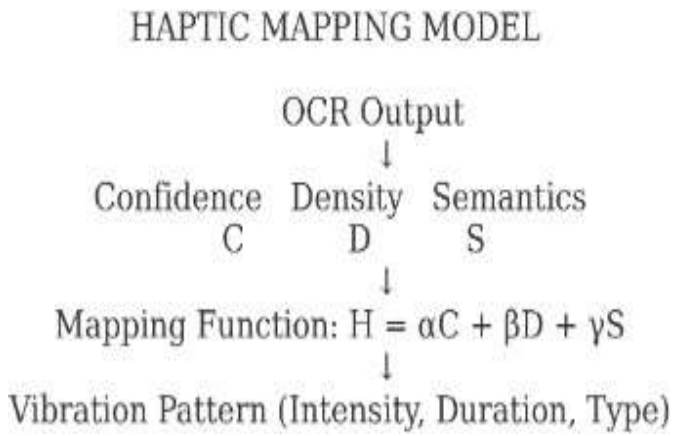

Figure 2. Conceptual diagram of the haptic-generation mechanism used to translate OCR-derived confidence, density, and semantic cues into vibration signals.

V. EXPERIMENTAL RESULTS

Tests have shown that SmartReader consistently outperforms well-known general OCR tools in terms of accuracy, stability, and accessibility features. Such an ensemble approach, combined with adaptive preprocessing, results in far better text output: 94.6% word accuracy and 96.8% character accuracy, both higher than those attained by single engines. Moreover, refinement based on SBERT reduces recognition slips typical for cluttered or low-quality inputs.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141298

Table I. OCR Accuracy Comparison

OCR Engine	Word Accuracy (%)	Character Accuracy (%)	Avg Confidence
Tesseract	82.4	90.3	0.74
EasyOCR	78.1	88.9	0.71
Ensemble (SmartReader)	94.6	96.8	0.87

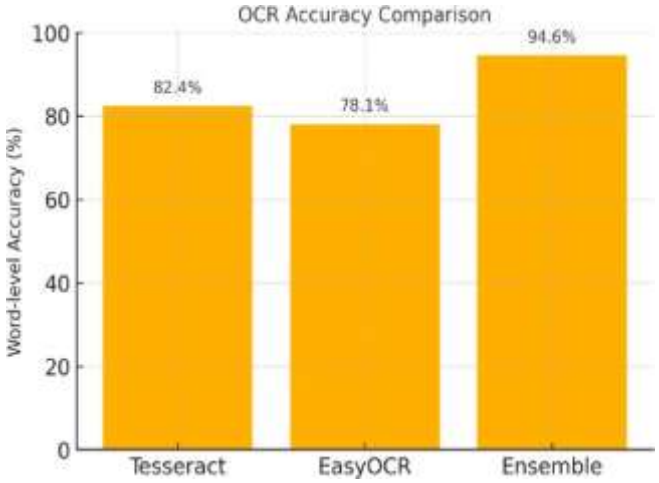

Figure 3. Comparative performance chart demonstrating how Tesseract, EasyOCR, and the combined ensemble approach differ in recognition accuracy and confidence.

Another salient feature in the behavior of SmartReader is that, although several components run during live scanning, the system can maintain an average end-to-end processing time per frame at about 217 ms to enable smooth interaction even in purely CPU-based environments. The integrated object analysis module furnishes contextual cues which standard OCR applications cannot, especially for documents that have been captured at imperfect angles or under varied lighting conditions.

Table II. Latency Breakdown

Module	Avg Latency (ms)
Preprocessing	32
OCR	118
YOLOv8	27
Analytics	40
Total	**217 ms**

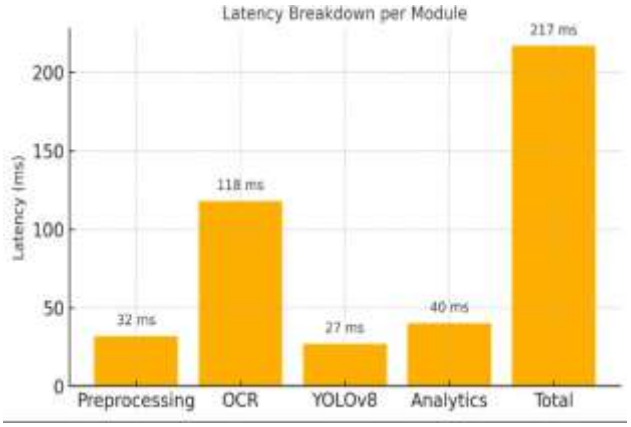

Figure 4. Breakdown of processing delays contributed by each functional block, highlighting the individual and cumulative latency of the framework

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141298

Beyond numerical enhancements, other modules, such as haptic signaling and deeper text analytics, introduce layers of support seldom found in conventional OCR utilities. Combined, they offer a much more adaptive reading tool, especially for those users who rely on non-visual channels to provide structure, emphasis, and navigation flow.

Table III. Object Detection Metrics

Metric	Value
mAP50	0.79
Precision	0.82
Recall	0.76

Table IV. Comparative Performance Analysis

Feature / Metric	Tesseract	EasyOCR	Mobile OCR Apps	SmartReader
Word Accuracy (%)	82.4	78.1	80–85	94.6
Character Accuracy (%)	90.3	88.9	88–92	96.8
Average Confidence	0.74	0.71	0.70–0.78	0.87
Average Latency (ms)	164	152	300–450	217
Preprocessing Quality	Medium	Medium	Low	High (multi-variant)
Semantic Correction	No	No	No	Yes (SBERT)
Object Detection Support	No	No	Partial	Yes (YOLOv8)
Real-Time Scanning	No	No	Limited	Yes (200 ms/frame)
Haptic Guidance	No	No	No	Yes
Text Analytics	No	No	Basic	Advanced
TTS Integration	Optional	Optional	Yes	Yes

Another salient feature in the behaviour of SmartReader is that, although several components run during live scanning, the system can maintain an average end-to-end processing time per frame at about 217 ms to enable smooth interaction even in purely CPU-based environments. The integrated object analysis module furnishes contextual cues which standard OCR applications cannot, especially for documents that have been captured at imperfect angles or under varied lighting conditions.

Beyond numerical enhancements, other modules, such as haptic signaling and deeper text analytics, introduce layers of support seldom found in conventional OCR utilities. Combined, they offer a much more adaptive reading tool, especially for those users who rely on non-visual channels to provide structure, emphasis, and navigation flow.

VI. CONCLUSION

Smart Reader demonstrates that judicious combination of multiple modes of perception can significantly enhance the access to written information by visually impaired users. Instead of mere audio output, the system merges text recognition, contextual scene cues, semantic refinement, and tactile signals to convey a reading flow that is significantly more informative. For the first time, this introduces a new pathway to comprehend text format and relative importance by touch that was previously unavailable to any other type of reader.

Results ranging from accuracy gains to stable performance using the standard CPU configurations indicate that this design is dependable and suitable for practical usage. Overall, the system provides a coherent way to transform visual content into interpretable, non-visual cues, hence allowing users to perceive the material in a more comprehensible and structured way. Herein, Smart Reader creates a valued alternative for reading print and digital text without relying on their vision.

ACKNOWLEDGMENT

The authors would like to convey their heartfelt gratitude to **Mr. Mohammed Shafiulla**, whose insights and valuable inputs throughout this project have been of immense help. The authors also thank **Dr. R. N. Kulkarni,** Head of the Department, Computer Science and Engineering, for providing the necessary resources and guidance. The support of the teaching and non-teaching staff of the Department of Computer Science and Engineering, Ballari Institute of Technology and Management, is gratefully acknowledged.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025
DOI: 10.17148/IJARCCE.2025.141298

REFERENCES

[1]. R. Smith, "An overview of the Tesseract OCR engine," in *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*, 2007, pp. 629–633.
[2]. J. Baek, G. Kim, J. Lee, and S. Yun, "What is wrong with scene text recognition model comparisons?" in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2019.
[3]. A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv:2004.10934, 2020.
[4]. S. Khusro, S. Rauf, M. Idrees, and M. A. Ullah, "Haptic feedback to assist blind people in indoor navigation: A survey," *Sensors*, vol. 22, no. 2, 2022.
[5]. C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. Int. Conf. Web and Social Media (ICWSM)*, 2014.
[6]. S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from text," in *Text Mining: Applications and Theory*, M. W. Berry and J. Kogan, Eds. Wiley, 2010.
[7]. M. Hersh, *Assistive Technology for Blind and Vision-Impaired People*. Springer, 2008.
[8]. B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.
[9]. J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv:1804.02767, 2018.
[10]. M. Grootendorst, "KeyBERT: Minimal keyword extraction with BERT," arXiv:2004.13699, 2020.

