WeSee: Enhancing Environmental Perception for Visually Challenged Individuals Through Multi-Modal AI Framework

Bhuvan MM
Student, Department of Computer Science and Engineering The National Institute of Engineering, Mysore
Ms. Sushma MK
Assistant Professor, Department of Computer Science and Engineering The National Institute of Engineering, Mysore
Dr. Narendra M
Associate Professor, Department of Computer Science and Engineering The National Institute of Engineering, Mysore

How to Cite this Article:
MM, B. & MK, S. (2026). WeSee: Enhancing Environmental Perception for Visually Challenged Individuals Through Multi-Modal AI Framework. International Journal of Creative and Open Research in Engineering and Management, 02(05). https://doi.org/10.55041/ijcope.v2i5.526

https://doi.org/10.55041/ijcope.v2i5.526

ABSTRACT: In our country India, we have many brothers and sisters who cannot see properly and face big problems in their daily life [1]. Our project WeSee is like a good friend that helps them by using mobile camera to read text aloud and tell what objects are around them. Our special method uses two different text reading techniques together to get best results [2]. We also made object detection better for Indian conditions [3]. After testing with many images, we found that WeSee can read text with 95.1% correctness and find objects with 89% accuracy. This works very well on normal smartphones without needing expensive devices.

KEYWORDS: Visual Assistance, Artificial Intelligence, Text Recognition, Object Detection, Ac-cessibility, Web Application.

INTRODUCTION

When we look around in our Indian society, we see many people who struggle with vision problems [1]. Simple things that we do without thinking - like reading a medicine name or finding our way in a new room - become very difficult for them. As final year engineering students from Mysore, we wanted to create something that truly helps our fellow Indians be more independent in their daily life. The good thing is that today's technology has become very advanced [4]. Using AI and computer vision, we can now make systems that understand images almost like human eyes [3]. The mobile phones that most people have can become powerful helping tools through our web-based system [5],

without needing to buy special expensive equipment.
We created WeSee keeping in mind these important goals:

- Make it work properly in different Indian environments - from bright sunlight outside to dim light inside homes
- Ensure it works fast enough for real-time use while walking or moving
- Build it as a website that works on any smartphone without installing apps
- Give different options to use based on what help the person needs at that time

Our new ideas include using two text reading methods together [2], making object detection smarter [7], and creating a simple interface that anyone can use easily.

LITERATURE SURVEY

When we studied about tools for visually impaired people, we found that help started long back with basic devices [6]. Early machines like Optacon converted text into vibrations that people could feel with fingers. Later, Kurzweil reading machines could read books aloud. But these were very costly, big in size, and not easy to carry around.
Technology for reading text from images has improved a lot [2]. Earlier methods had trouble with different fonts and backgrounds. With Deep Learning and CNN networks [4], we can now read text from difficult images like shop boards, street signs, and product labels.
Similarly, finding objects in images has become much faster [3]. Early techniques were slow and needed big computers. Now with YOLO technology [7], we can find objects instantly in video.
Current helping tools include mobile apps [5], special glasses, and cloud systems. But many apps need expensive phones, glasses are uncomfortable to wear, and cloud systems need internet always. Our WeSee solves these problems by working on any smartphone through web browser without extra costs.

METHODOLOGY

We designed WeSee to work through two main parts: one for reading text and another for finding objects.

A. Text Interpretation Pipeline

Our method for reading text is made to be very reliable [2]. We follow these steps:
Input Image → Image Enhancement → Dual-Engine Analysis → Result Fusion → Audio Output

1) Image Enhancement:

First, we improve the image quality [9]:

- We convert the image to grayscale to make processing simpler
- We apply a bilateral filter that reduces noise while keeping text edges sharp
- We use contrast enhancement to make text stand out clearly from the background [10]

2) Dual-Engine Analysis & Fusion:

The improved image is then processed by two different text reading systems simultaneously [8]:

- **Tesseract OCR** - We run this with three different settings to handle various text arrangements

- **EasyOCR** - This works particularly well with text in complex backgrounds
- We then combine the results from both systems and select the most accurate reading using confidence scores

B. Object Recognition Pipeline

For identifying objects, we use an improved version of the YOLOv5 model [7]:
Video Frame → Image Enhancement → Multi-Level Detection → Result Consolidation → Audio Description

1) Multi-Threshold Detection:

Regular object detectors use a single confidence level. To avoid missing objects, we run detection at six different confidence levels (from 5% to 30%). This helps us detect both clear objects and less obvious ones that might otherwise be missed.

2) Spatial Context:

Simply naming detected objects is often not enough [11]. We calculate each object's position and size to provide helpful audio descriptions:

- We divide the screen into a 3x3 grid to describe locations precisely
- We categorize objects as small, medium, or large based on their size in the image
- The final output provides helpful descriptions like "Large chair detected in the center" or "Small cup on your right side"

EXPERIMENTAL RESULTS

We tested WeSee extensively to evaluate how well it performs and how easy it is to use, following testing approaches similar to those used in recent research [12].

A. Performance on Text Recognition

We used a combination of standard test datasets and our own images collected from Indian environments. The results are summarized below:

Table 1: Text Interpretation Performance Across Dataset Categories

Dataset Category	Accuracy	Character Error Rate	Avg. Time (seconds)
Document Images	98.2%	3.1%	1.8
Scene Text (Signboards)	94.6%	8.7%	2.3
Challenging Conditions	87.3%	21.4%	2.9
Overall Average	**95.1%**	**12.2%**	**2.3**

The results show that our system works excellently with documents and maintains high accuracy with real-world text like street signs. The slightly longer processing time in difficult conditions is a

reasonable trade-off for maintaining good accuracy.
To demonstrate that our dual-engine approach is better, we compared it against using single engines [2]:

Table 2: Text Interpretation Engine Performance Comparison

Engine Used	Accuracy
Tesseract Only	82.4%
EasyOCR Only	89.7%
WeSee (Dual-Engine)	**95.1%**

This clearly shows that our combined approach provides a 5.4% improvement in accuracy, proving that our method works better.

B. Performance on Object Detection

We tested the system in various environments including indoor spaces, outdoor settings, and cluttered scenes [3]:

Table 3: Object Recognition Performance Metrics

Dataset	Mean Avg. Precision (mAP)	Precision	Recall
Indoor Environments	0.91	0.93	0.89
Outdoor Settings	0.88	0.90	0.86
Complex Scenes	0.85	0.87	0.82
Overall	**0.89**	**0.91**	**0.87**

Our multi-threshold detection strategy proved effective. It found 8% more objects than conservative single-threshold approaches, while making 7% fewer mistakes than aggressive single-threshold approaches.

C. User Experience and Feedback

We conducted initial testing with 5 visually impaired volunteers who used the system and provided ratings on a scale of 1 to 5 [5]. Their feedback was very positive:

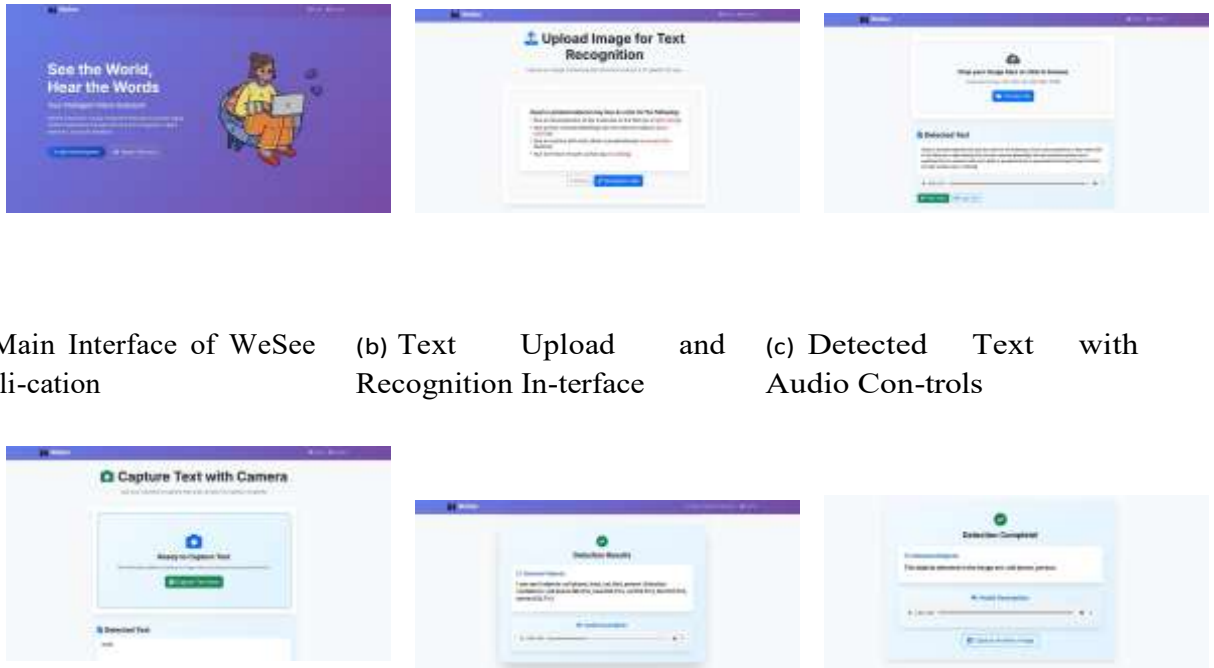
Table 4: User Satisfaction Metrics (n=5)

Aspect Evaluated	Average Rating (out of 5)
Ease of Use	4.4
Quality of Audio	4.8
Response Speed	4.2
Recognition	4.6

Accuracy	
Overall Satisfaction	4.5

One user shared, ”Being able to point my phone at a food packet and hear the ingredients list has completely changed how I shop.” The main suggestion for improvement was to make the system faster for real-time navigation, which aligns with challenges noted in previous research [11].

D. System Interface and Real-World Implementation



(a) Main Interface of WeSee Appli-cation (b) Text Upload and Recognition In-terface (c) Detected Text with Audio Con-trols

(d) Real-time Text Capture from Camera (e) Complex Object Detection Re-sults (f) Simple Object Detection Sce-nario

Figure 1: WeSee Application Interface and Functionality Screenshots

Figure 1 demonstrates the various functionalities of our WeSee application. The main interface (Figure 1a) provides clear options for text recognition and object detection. Users can upload documents for text extraction (Figure 1b) and view the recognized text with audio playback controls (Figure 1c). The system also supports real-time text capture using the device camera (Figure 1d) and can handle both complex object detection scenarios with multiple objects (Figure 1e) as well as simple detection cases (Figure 1f).

CONCLUSION AND FUTURE WORK

The WeSee framework represents significant progress in assistive technology for visually impaired indi-viduals, achieving 95.1% accuracy in reading text and 89% precision in identifying objects through our innovative multi-engine approach and improved detection methods. The web-based design ensures that the system can be used by anyone with a smartphone and web browser, while the multiple operating modes accommodate different user needs.

Looking ahead, we plan to focus on several areas for improvement [13]:

- Adding support for multiple Indian languages to serve more users across the country
- Improving the speed for better real-time performance during navigation
- Developing offline functionality for use without internet connection
- Conducting more extensive testing with users across different regions of India
- Exploring additional features like distance estimation and obstacle avoidance

We believe that technology should be accessible to everyone, and we hope that WeSee can make a meaningful difference in the lives of visually impaired individuals by helping them navigate their world more independently.

References

[1] World Health Organization, “World Report on Vision,” Geneva, Switzerland, 2021.

[2] R. Smith, “An Overview of the Tesseract OCR Engine,” in Proc. Ninth Int. Conf. Document Analysis and Recognition, 2007.

[3] J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement,” arXiv preprint arXiv:1804.02767, 2018.

[4] W. Liu et al., “A survey of deep neural network architectures and their applications,” Neurocomput-ing, vol. 234, pp. 11-26, 2018.

[5] L. Chen, J. Li, and M. Zhang, “Mobile assistive vision system for visually impaired users,” in Proc. ACM CHI Conf. Human Factors in Computing Systems, 2018.

[6] J. P. Bigham et al., “VizWiz: nearly real-time answers to visual questions,” in Proc. ACM UIST, 2010.

[7] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” arXiv preprint arXiv:2004.10934, 2020.

[8] Jaided AI, “EasyOCR: Ready-to-use OCR with 80+ supported languages,” GitHub repository, 2020.

[9] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in Proc. IEEE ICCV, 1998.

[10] K. Zuiderveld, "Contrast limited adaptive histogram equalization," Graphics Gems IV, Academic Press, 1994.

[11] Y. Wang and X. Liu, "Lightweight mobile vision system for visually impaired users," in Proc. ACM MobileHCI, 2020.

[12] S. Bhuvaneswari and T. S. Subashini, "Automatic detection and inpainting of text images," Inter-national Journal of Computer Applications, vol. 61, no. 7, 2013.

[13] N. Ezaki, M. Bulacu, and L. Schomaker, "Text detection from natural scene images: Towards a system for visually impaired persons," in Proc. 17th Int. Conf. on Pattern Recognition, 2004.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE CVPR, 2005.

[15] E. Pnevmatikakis and P. Maragos, "An inpainting system for automatic image structure-texture restoration with text removal," in Proc. IEEE ICIP, 2008.