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 20, April 2026
Impact Factor: 8.2

# Real-Time Scene Understanding for the Visually Impaired using Image Captioning and Audio Feedback

**Prof. S. B. Dhadake[1], Ketan Bhoskar[2], Rajat Pathak[3], Malhar Shinde[4], Piyush Landge[5]**
Assistant Professor, Smt. Kashibai Navale College of Engineering, Pune, India[1]
Undergraduate Students, Smt. Kashibai Navale College of Engineering, Pune, India[2-5]

**Abstract:** *People with visual impairments run into all sorts of problems when they try to move around on their own. Without sight, it's tough to know what's going on around you. That's where this project comes in. We've built a deep learning system that gives real-time scene descriptions through image captioning and audio feedback. Here's how it works: a camera—either on a cane or wearable device—grabs images of the environment. The system uses YOLO object detection to spot things like obstacles or important features. Then, a CNN-LSTM model turns those details into full sentences, explaining what's in the scene. Finally, a text-to-speech module reads those sentences out loud, so users get instant feedback about their surroundings. By combining computer vision and natural language processing, this system helps people with visual impairments become more aware of their space and move around more independently. Tests show it runs in real time, stays accurate, and understands context, so it's a strong option for accessible navigation and support*

**Keywords:** Image Captioning, YOLO, CNN-LSTM, Text-to-Speech, Assistive Technology, Blind Cane Navigation.

## I. INTRODUCTION

Most of what we know about the world comes through our eyes. But for people with visual impairments, just getting around—or even figuring out what's happening nearby—can be a real struggle. Lately, though, AI has taken some big leaps, especially in computer vision and language. Now, it's possible to turn what's in front of someone into detailed, real-time descriptions they can hear or read [1], [2]. This project is all about making that happen. The idea is simple: grab live images from a person's surroundings, spot important objects and get a sense of the scene and then turn all that into speech [3], [4]. The tech behind it uses deep learning—stuff like YOLO for finding objects, and either CNN-LSTM or Transformer models to generate more natural, story-like captions. Instead of just rattling off object names ("chair, table, lamp"), the system can say things like, "There's a chair next to a small table with a lamp on it [5]. That kind of detail really helps people understand what's going on, making it easier to move around safely and confidently. We're focused on keeping everything fast, lightweight, and easy to use [1], [9]. You could run this on a phone, a pair of smart glasses, or even a high-tech cane [6]. In the end, it's about more than just accessibility—it's about giving visually impaired people more freedom and helping them feel included in everyday life [7], [8].

## II. LITERATURE SURVEY

| Sr.no | Year | Author | Description |
|---|---|---|---|
| 1. | 2024–2025 | Gao at al. / DistinctAD (2024-25) | Recent advances in automated audio description (AD) generation leveraging VLMs/LLMs — includes DistinctAD (two-stage AD generation) and NAACL findings on AD generation in the era 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 20, April 2026
Impact Factor: 8.2

| | | | |
|---|---|---|---|
| | | | large multimodal models.<br>These works are important when integrating natural-sounding audio outputs with captioning systems. |
| 2. | 2024 | University of Michigan team | Prototype tools that use generative AI (VLMs + LLMs) to produce real-time text + audio descriptions from camera images for blind users — demonstrates feasibility of low-latency caption + TTS pipelines. |
| 3. | 2024 | Yuan et al. (DSC-Net) | DSC-Net: semantic segmentation tailored to "blind-road" scenarios — high mIoU on specialized datasets; useful for lane/sidewalk/obstacle parsing in assistive navigation systems. |
| 4. | 2024 | Abidi et al. | Comprehensive review (2024) of navigation systems for visually impaired — summarizes recent advances in sensors, deep-learning detection/segmentation, and multimodal feedback for mobility assistance. |
| 5. | 2023 | Ma et al. (EOS) | EOS — a wearable system performing efficient obstacle segmentation and real-time guidance for blind users; demonstrates fast segmentation + haptic/audio feedback for safe walking. |
| 6. | 2023 | Khadidja Delloul / Slimane Larabi | Research on egocentric (wearable camera) scene description for blind and visually impaired users — explores user needs for turn-by-turn, context-aware descriptions during navigation |
| 7. | 2022 | Tiwary et al. | Deep-learning approach for automated image captions to assist blind users in e-commerce (food/grocery item identification) shows applied captioning pipelines + TTS for accessibility tasks |
| 8. | 2022 | Kassem et al. | Real-time scene monitoring device for deaf-blind people combining mm-Wave sensing, tracking and non-visual feedback — relevant for real-time environmental monitoring and multimodal feedback design. |
| 9. | 2022 | Dognin et al. | "Image Captioning as an Assistive Technology" — work on improving image captioning for accessibility, lessons from captioning challenges (e.g., VizWiz) and engineering choices to make captions useful for blind users. |
| 10. | 2021 | (Survey) Messaoudi et al. | Review of navigation-assistive tools and technologies for visually impaired users — surveys sensors, computer-vision methods, wearable systems and the challenges in indoor/outdoor navigation for BLV users. |

## III. OVERVIEW OF THE SYSTEM

The proposed solution will help those individuals who are visually impaired by providing accurate real-time information about their environment via audio. This solution will capture images (or video frames) of a physical environment using a camera and use deep learning methods to process the information captured through the camera [1], [2].

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 20, April 2026
Impact Factor: 8.2

Object detection models such as the YOLO model detect and identify objects that exist within the scene, while CNN-based feature extraction generates visual features from those images. Image captioning models (CNN–LSTM, Vision-Language) will use these created features to produce a textual description of the visual data captured by the camera [3].
Once the textual description of the scene is generated, the description will be converted to audio using a Text-to-Speech engine so that the audio output will be communicated to the user via either headphones or speakers, allowing the visually impaired to understand their environment and providing further assistance in navigating within it [4], [6].
The combination of computer vision, natural language processing, and speech synthesis provides improved accessibility, independence, and overall situational awareness to the visually impaired individuals [7], [8].

**PROPOSED SYSTEM:**

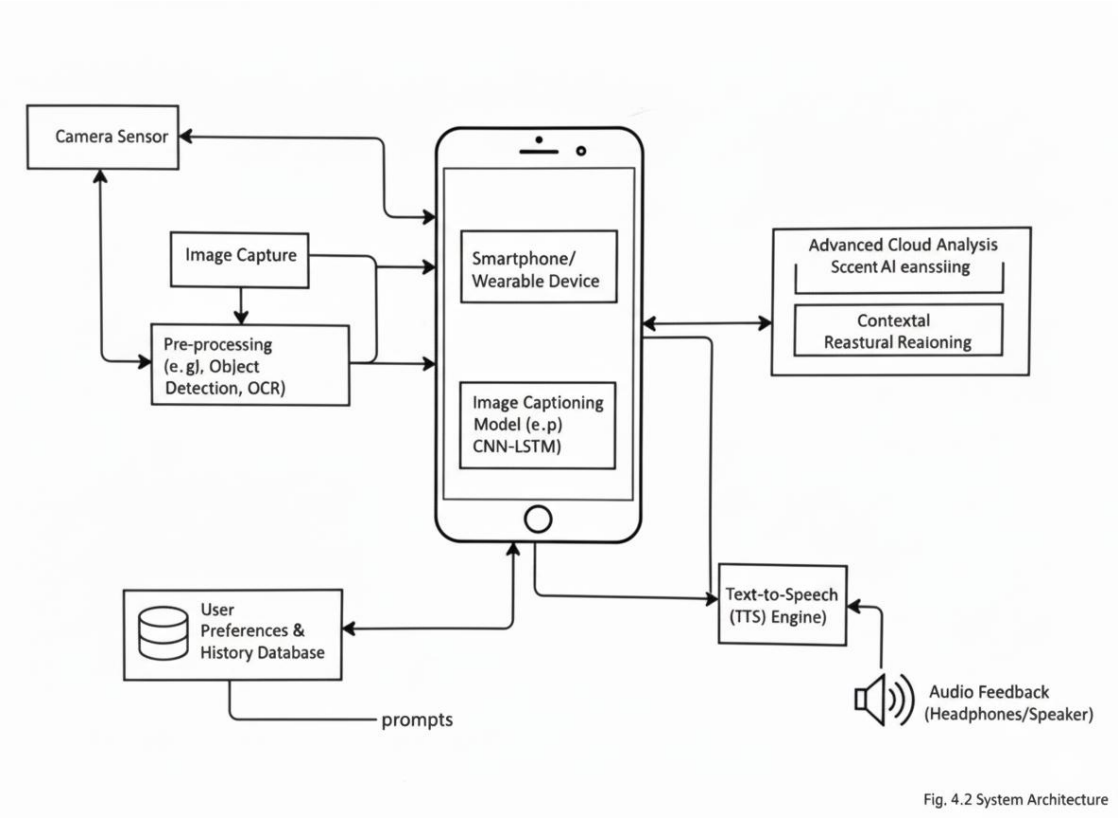



Fig: System Architecture

Using a smartphone and/or wearables device's camera (camera sensor), Real-time data (pictures) are captured (from the environment). These pictures then go through a pre-processing phase (object detection including identifying known objects as well as extracting features of the object) [6]. A deep CNN-LSTM combination uses these pre-processed images to perform image captioning (returns a text caption for the image) [3]. If necessary a cloud-based AI will conduct additional analysis to provide context for all AI analyses therefore, improving captioning accuracy with AI reasoning through Cloud-sourced images; with the AI-based text captions found, a TTS will output audio spoken via headphones and/or speakers [4], [7]; finally, a database of user preferences (previously selected or chosen by the user for their interaction with the device) is used to provide the user with consistent response patterns tailored specifically to each individual user's responses/behavior in ordering or having an item delivered and/or how they choose to respond [1].

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 20, April 2026
Impact Factor: 8.2

**Methodology:**

The proposed system will capture real-time images through a camera and process them through a series of deep learning models. The image will be preprocessed for better quality [2]. A YOLO-based object detection model will detect the objects present in the scene. A Convolutional Neural Network (CNN) will be used to detect the crucial features present in the image [3].

These features will be further processed by an image captioning model (CNN-LSTM or Vision Language Model), which will generate a sentence describing the scene [3], [4]. The sentence can be further processed by a Large Language Model (LLM) for better clarity and context [6]. Finally, the sentence will be converted to speech by a Text-to-Speech (TTS) engine [1], [10].

This methodology can provide efficient computer vision and natural language processing-based assistance for the visually impaired.

## IV. ALGORITHMS

The proposed algorithm aims to provide real-time scene understanding for visually impaired users by converting visual input into meaningful audio descriptions [2]. The algorithm integrates object detection, image captioning, and text-to-speech synthesis in a sequential and optimized manner to ensure low latency and high accuracy [8].

Step-by-Step Description

**System Initialization:**

The system is initialized when the user activates the application using a button or voice command. User preferences such as language, speech rate, and volume are loaded from the user database.

**Image/Video Capture:**

A camera mounted on a wearable device continuously captures images or video frames of the surrounding environment in real time.

**Pre-processing:**

Each captured frame is resized, normalized, and enhanced to remove noise and improve visual quality. This step ensures that the input image is suitable for deep learning models.

**Real-Time Object Detection (YOLO):**

The pre-processed frame is passed to the YOLO object detection model, which detects multiple objects simultaneously. The model outputs object labels, bounding box coordinates, and confidence scores.

**Feature Extraction (CNN):**

A convolutional neural network extracts high-level visual features from the image and detected objects. These features represent the semantic content of the scene.

**Caption Generation (CNN–LSTM / VLM):**

The extracted features and detected object information are fed into the caption generation module. A CNN–LSTM or Vision-Language Model generates a coherent and context-aware textual description of the scene.

**Text-to-Speech Conversion (TTS):**

The final caption is converted into natural-sounding speech using a text-to-speech engine. User preferences are applied to control voice type and speaking speed.

**Audio Feedback Delivery:**

The synthesized audio is delivered to the user through headphones or a speaker, providing real-time awareness of the surroundings.

**Logging and Personalization:**

Detected objects, captions, and user feedback are stored in the database for future analysis and system improvement.

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 20, April 2026
Impact Factor: 8.2

## V. FUTURE SCOPE

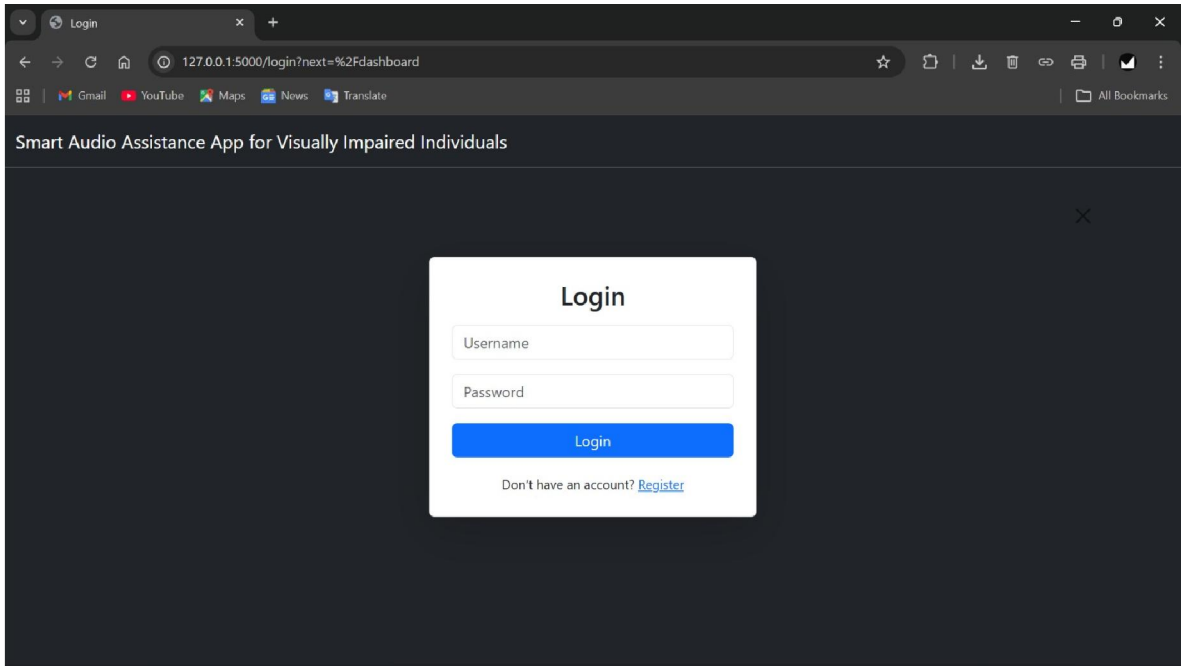
Further potential improvements could come from incorporating significant advancements in artificial intelligence methods, as well as in hardware technologies. Future developments could include enhancing object class recognition accuracy and generating scene captions by utilizing Enhanced Vision–Language Models (VLMs) or Large Language Models (LLMs) to provide additional contextually relevant and detailed descriptions of complex visual environments [6], [7].

Implementing GPS systems, combined with navigational systems, will allow an additional feature for visually impaired users to receive location-based assistance together with an accurate description of the environment [8]. Wireless products such as smart glasses or smart canes could enable the system to be more portable and easier to use every day [6].

Adding multilingual capabilities to the smart assistive technology, along with customized voice commands, would enable users from different regions to utilize the system more easily. Continued research may also examine adding lightweight artificial intelligence (AI) methods to utilize edge computing to reduce processing time and minimize reliance on cloud server-based technologies. All these improvements will help increase the reliability, efficiency, and usability of assistive technologies in practical applications [5], [10].

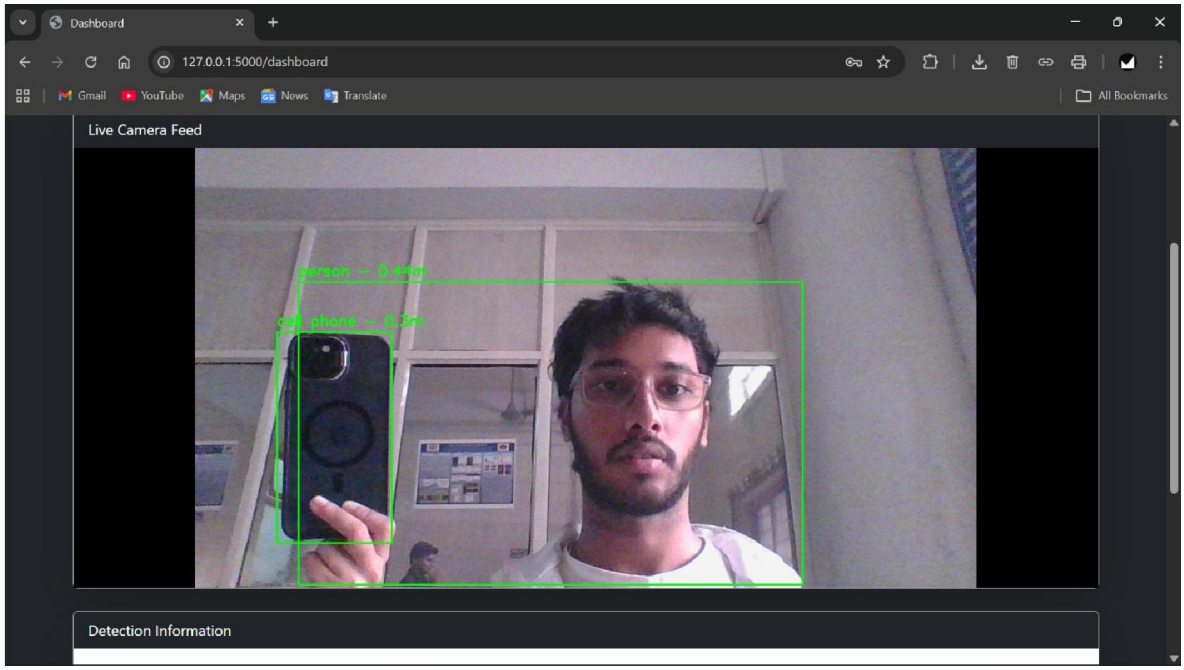
## VI. RESULT

Login Page:



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 20, April 2026
Impact Factor: 8.2

Result Page:




## VII. CONCLUSION

There have been many advancements within assistive technologies using computer vision and deep learning techniques over the last few years. Methods using object detection, image captioning, as well as text-to-speech systems allow computers to analyze visual scenes to create meaningful audio descriptions for people with vision impairment to understand their environment, promoting greater independence and safety while navigating [1], [2].

Technology is constantly evolving and improving using various techniques for analyzing and understanding how we visually interpret visual scenes and objects. Approaches such as CNN-Based feature extraction, YOLO for object detection, and image captioning through CNN-LSTM and Visual-Language Models have provided improved accuracy and effectiveness [3], [4]; however, there are still challenges with respect to system reliability, providing real-time responses, and providing contextual understanding of the information being processed [7].

Future research in assistive technology using intelligent scene understanding systems should focus on developing more efficient multimodal models, providing greater contextual reasoning, and optimizing the systems for use with wearable devices [6], [8]. Continuing advancements in artificial intelligence and assistive technology will enable intelligent scene understanding systems to enhance accessibility and improve the quality of life for individuals with visual impairments [1], [9].

## REFERENCES

[1] Arystanbekov, B., et al., “Image Captioning for the Visually Impaired and Blind.” IEEE/Medical AI Research, 2023. This work combines image captioning with text-to-speech to provide assistive descriptions of surroundings for visually impaired users. (Batyr Arystanbekov, 2023)

[2] Sharma, D., “Evolution of Visual Data Captioning: Methods, Datasets and Applications.” Artificial Intelligence Review, 2023. The paper reviews modern image captioning techniques and highlights applications such as assistive technology and scene understanding. (Dhruv Sharma, 2023)

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 20, April 2026
Impact Factor: 8.2

[3] Farkh, R., et al., "Image Captioning Using Multimodal Deep Learning." Computer Modeling in Engineering & Sciences, 2024. The authors integrate YOLO-based object detection with deep learning models for improved image caption generation. (Rihem Farkh, 2024)
[4] Kavitha, P. V., "Image Captioning Deep Learning Model Using ResNet50 and Hybrid LSTM-GRU." International Journal of Computer Vision Applications, 2025. This research proposes an encoder–decoder architecture for generating meaningful captions from images. (P. V. Kavitha, 2025)
[5] Khan, A., "Novel Image Captioning Technique Using Deep Learning." International Conference on Knowledge Computing, 2025. The study uses CNN and RNN architectures to generate context-aware captions for visual scenes. (Abdullah Khan, 2025)
[6] Parseh, M. J., "Graph-Based Image Captioning with Semantic and Spatial Relationships." Information Processing Systems, 2025. This work improves caption generation by modelling relationships between objects in images. (Mohammad Javad Parseh, 2025)
[7] Thobhani, A., "A Survey on Enhancing Image Captioning with Advanced Deep Learning Methods." Computational Methods in Engineering, 2025. This survey discusses transformer-based captioning models and future research directions in multimodal AI. (Alaa Thobhani, 2025)
[8] Bhosale, C. S., "A Review of Image Captioning Techniques and Deep Learning Approaches." Artificial Intelligence Research Journal, 2025. The paper reviews retrieval-based, template-based, and deep learning captioning methods. (Chaitanya S. Bhosale, 2025)
[9] "Deep Learning-Based Image Captioning for Visually Impaired People." ResearchGate Publication, 2023. This study focuses on developing captioning models that help blind individuals understand surrounding scenes. (R. Kavitha, 2023)
[10] "Speech-Based Real-World Scene Understanding for Assistive Care of the Visually Impaired." AI Assistive Systems Research, 2025. The system integrates image captioning and speech synthesis for real-time environmental understanding. (Tarun Sunil, 2025)