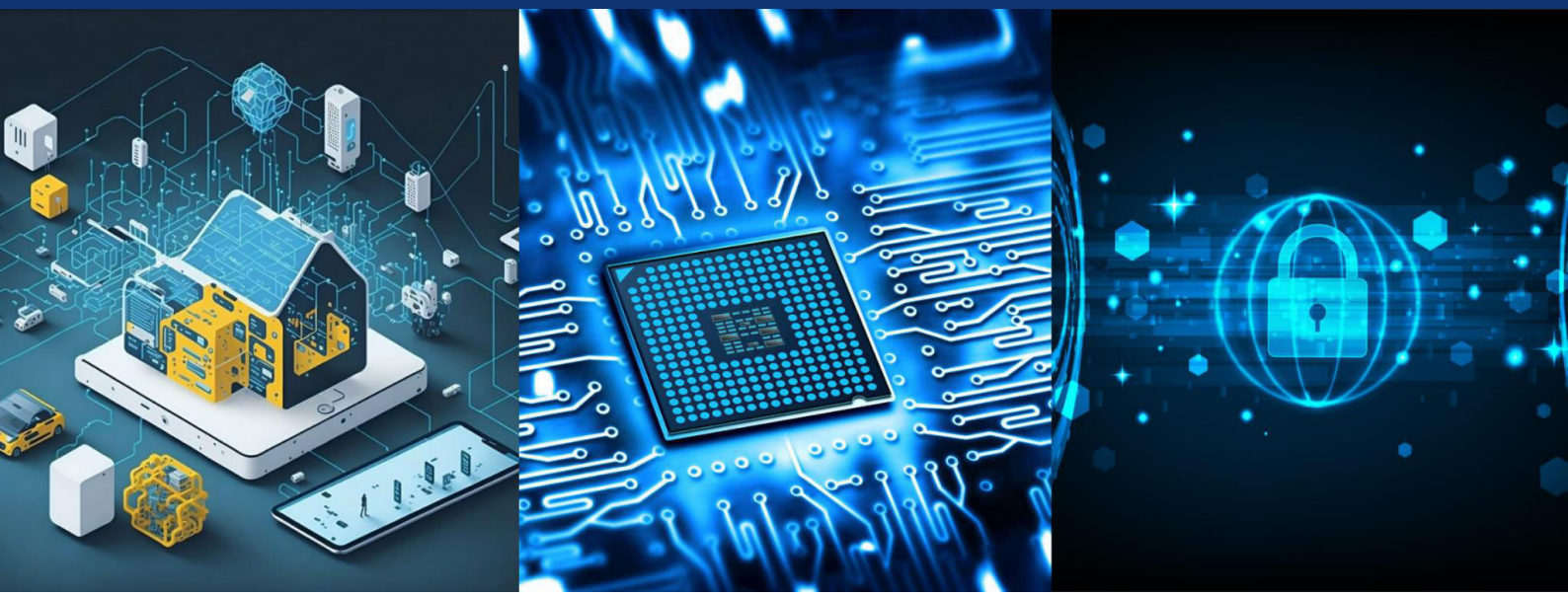
IJIRCCE

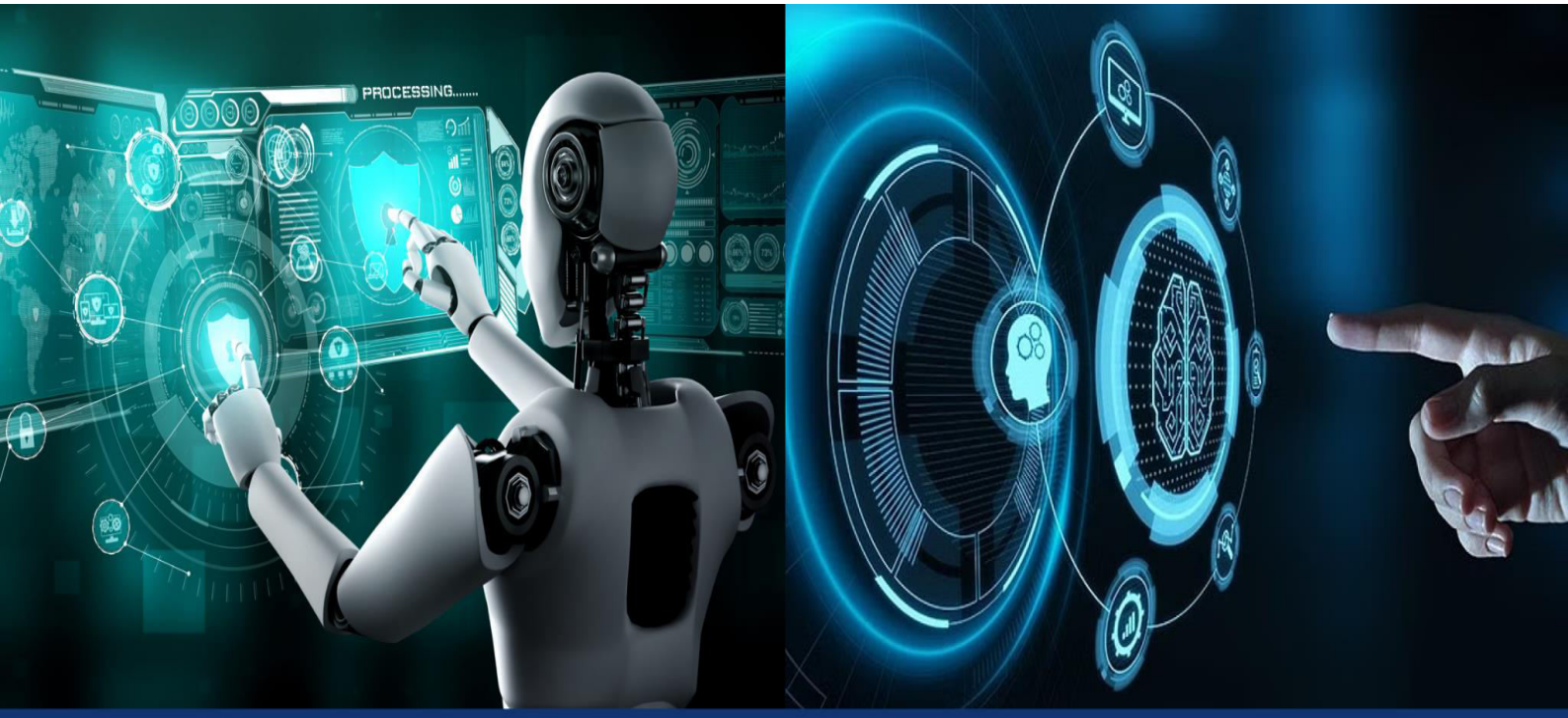
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 5, May 2025

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305078

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VisionAID: AI-Powered Image Caption Generator for Visually Impaired

Samrudh BS[1], Renukaprasad VG[2], Inchara Patel MS[3], Ranjitha HG[4], Prof. Rachana G Sunkad[5]

B.E Student, Dept. of CSE, BIET, DVG, Karnataka, India[1]

B.E Student, Dept. of CSE, BIET, DVG, Karnataka, India[2]

B.E Student, Dept. of CSE, BIET, DVG, Karnataka, India[3]

B.E Student, Dept. of CSE, BIET, DVG, Karnataka, India[4]

Assistant Professor, Dept. of CSE, BIET, DVG, Karnataka, India[5]

ABSTRACT: Vision- AID is an innovative AI- powered image captioning system developed to empower visually bloodied individualities by transubstantiating visual data into meaningful audio descriptions. using the power of deep literacy, the system uses apre-trained InceptionV3 convolutional neural network to prize high- position visual features from input images. These features are also fed into an LSTM- grounded decoder model trained on the Flickr8k dataset to induce grammatically coherent and contextually accurate captions. To insure availability, the generated captions are vocalized using the pyttsx3 textbook- to- speech machine, allowing druggies to hear descriptions of images in real time. The result runs efficiently on original systems and is erected using Python with TensorFlow and Keras for the modeling factors. VisionAID islands the gap between vision and understanding, making digital content and surroundings more accessible and inclusive for the visually disabled community.

KEYWORDS: VisionAID, Image Captioning, YOLOv5, BLIP Model, Text-to-Speech, CNN,NLP

I. INTRODUCTION

VisionAID is a system powered by AI and designed to increase visually impaired persons' independence and situational understanding. It integrates cutting-edge technologies like image captioning, object detection, and voice navigation into an affordable, multi-tool assistant. Leveraging deep learning models like BLIP for image captioning and YOLOv5 for real-time object recognition, VisionAID is capable of processing uploaded or captured images, creating descriptive captions, and speechifying them. This enables them to learn quickly about what's around them independently, and even common tasks and activities are more manageable and achievable.

A particular highlight of VisionAID is its live object detection, which functions through a live webcam connection. The system identifies objects around the user in their environment and speaks out their names, enabling the user to navigate around obstructions and stay aware in changing environments. In addition, the smart navigation assistant offers voice guidance with the use of Google Maps API. Users can enter destinations with voice input, and the system reports back with spoken step-by-step directions that are supported by manual as well as GPS-guided navigation.

Developed on Python, Flask, OpenCV, and JavaScript, VisionAID is made lightweight, real-time, and user-friendly for use across devices. It has a simple and elegant interface with big buttons and voice interaction, making it accessible even to users who are not familiar with reading or typing. With offline capabilities and a single platform for several assistive functions, VisionAID overcomes the major shortcomings of conventional tools and current apps, providing a more inclusive and scalable solution for the 285 million visually impaired individuals globally.

II. RELATED WORK

Chen and Kumar (2025) outline a revolutionary multimodal AI model that is tailored to support visually impaired people's navigation and perception of the environment. The model combines different artificial intelligence modalities, such as computer vision, natural language processing, and sound feedback systems, to develop a holistic solution that supports mobility and independence.

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305078

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

At the center of the framework is data collection in real-time from cameras and sensors that enable the system to effectively understand the user's environment. With visual inputs being analyzed, the AI can pick out obstacles, landmarks, and other notable points in the environment. The system then converts all this into audio descriptions that will guide users through their environment in a safe way.

The multimodal design is advantageous because it accommodates various user needs and preferences. Some users might find it suitable to receive information in the form of auditory descriptions, for example, while others could be content with haptic feedback or visual cues on a mobile phone. Offering information in more than one format makes the framework more accessible and usable, allowing visually impaired users to explore unfamiliar environments with increased confidence and independence.

Tanaka and Singh (2025) are concerned with a particular use of AI for the aid of visually impaired persons via scene detection. Their work uses the YOLOv5 (You Only Look Once version 5) object detection model to design a system capable of detecting multiple scenes and objects in real-time.

The software is set to generate voice notifications that identify detected objects and scenes, promoting the situation awareness of visually impaired individuals. Through the exploitation of YOLOv5 speed and accuracy, the app can perform rapid environmental analysis and provide timely feedback in audio. The feature is important in assisting users in navigation safely and efficiently, as it helps them better comprehend their environment and make informed choices when accessing various environments.

Ibrahim and Garcia (2024) delve into using transformer models to create sophisticated captioning systems that are specifically optimized for assistive technology. In their research, they concentrate on real-time captioning of visual content so that it can be made more accessible to the visually impaired.

Utilizing the potential of transformers, the system can serve up context-specific and correct explanations of images or videos. It facilitates the overall user experience to an extent by enabling visually challenged users to view visual content which they might find impossible to comprehend otherwise. By making sure generated captions are meaningful in addition to being accurate, the transformer-based methodology helps enforce inclusiveness to a significant level while consuming digital content.

Mukherjee et al. (2024) present a mobile app that uses AI to translate scenes in real-time for visually impaired individuals. The app uses computer vision algorithms to examine the camera stream from a smartphone, detecting objects, individuals, and other useful features in the surroundings.

The interpretation is transferred to the user in audio form, which makes them able to navigate and communicate with their environment more effectively. The mobile technology focus ensures the solution is both convenient and available, with the user able to take the assistance tool with them wherever they move. The live scene interpretation further boosts the level of independence available to the visually impaired, granting them the instruments they require in order to fully participate in their environment.

Novak and Sharma (2023) introduce a context-aware captioning system that translates visual data into speech for accessibility. This system utilizes cutting-edge image recognition and natural language processing methods to interpret visual data and produce contextually appropriate spoken descriptions. By grasping the context of how an image or scene is displayed, the system is better able to give more substantive and helpful descriptions, enriching the experience for visually impaired users. The solution works to close the gap between visual media and accessibility so that those with visual impairments will be able to access content which would otherwise be impossible for them to do. The context-aware nature of the system allows for a richer understanding of visual information, making it a valuable tool for enhancing accessibility in various settings.

III. PROPOSED ALGORITHM

A. Design Considerations:

The system to be designed is an AI-powered Image Caption Generator to assist visually impaired individuals by creating descriptive, context-relevant, and real-time descriptions of images. It applies advanced deep learning and

natural language processing (NLP) techniques to generate meaningful captions, which are subsequently converted to speech using Text-to-Speech (TTS) technology. The architecture consists of key modules: an Image Processing Module with the use of Convolutional Neural Networks (CNNs) for feature extraction from images, and a Caption Generation Module using Transformer models like BLIP or GPT-4 Vision. These models make use of attention mechanisms to focus their attention on important regions of the image in order to generate accurate and descriptive captions.

To enhance accessibility and usability, the system also has Real-Time and Offline Capability using Edge AI techniques to work without internet. It also incorporates Advanced Scene Understanding to comprehend multi-object complex images and explain relationship and interaction within scenes. Translated text captions are converted to speech output by the TTS module, providing audio feedback through headphones, speakers, or intelligent devices. Extensive Usability and Safety Testing with blind users will make the system operational, user-friendly, and continually optimized based on actual feedback.

B. Description of the Algorithm:

The Image Caption Generator powered by AI is intended to support visually impaired people by creating live, detailed, and context-related audio descriptions of images. The system combines several AI technologies such as deep learning, computer vision, and natural language processing in six core modules: Image Processing, Caption Generation, Text-to-Speech Conversion, Real-Time and Offline Capability, Advanced Scene Understanding, and Evaluation & Testing.

1. Image Processing Module

This module employs Convolutional Neural Networks (CNNs) to obtain key visual features from input images. These are utilized by the system to detect objects, actions, and surroundings, which provide the basis for creating meaningful captions.

2. Caption Generation Module

A Transformer-based model like BLIP or GPT-4 Vision is employed to produce natural language descriptions. By using attention mechanisms, the model attends to the most descriptive areas of the image to produce more precise, context-aware captions.

3. Text-to-Speech (TTS) Module

The synthesized textual descriptions are translated into sound through TTS technology. It enables visually impaired users to listen to the image descriptions via headphones, speakers, or built-in smart devices, improving accessibility.

4. Real-Time and Offline Capability Module

Edge AI methods are utilized to make the system functional without the need for an internet connection, enabling real-time captioning even in low-connectivity contexts. This renders the system extremely pragmatic and adaptable to different real-world applications.

5. Advanced Scene Understanding Module

This module enhances the system's capability to understand complicated scenes through recognizing relationships and interactions between more than one object. It refines the depth and utility of the image captions, particularly in dynamic or cluttered settings.

IV. EVALUATION & TESTING MODULE

Extensive user testing with visually impaired users is done to evaluate system accuracy, usability, and accessibility. Iterative feedback-driven improvements are implemented. A large-scale set of performance maps tests aspects like answer correctness, context relevancy, applicability, and faithfulness in both text and image modalities, providing insight into the strengths and weaknesses of each model.

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305078

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

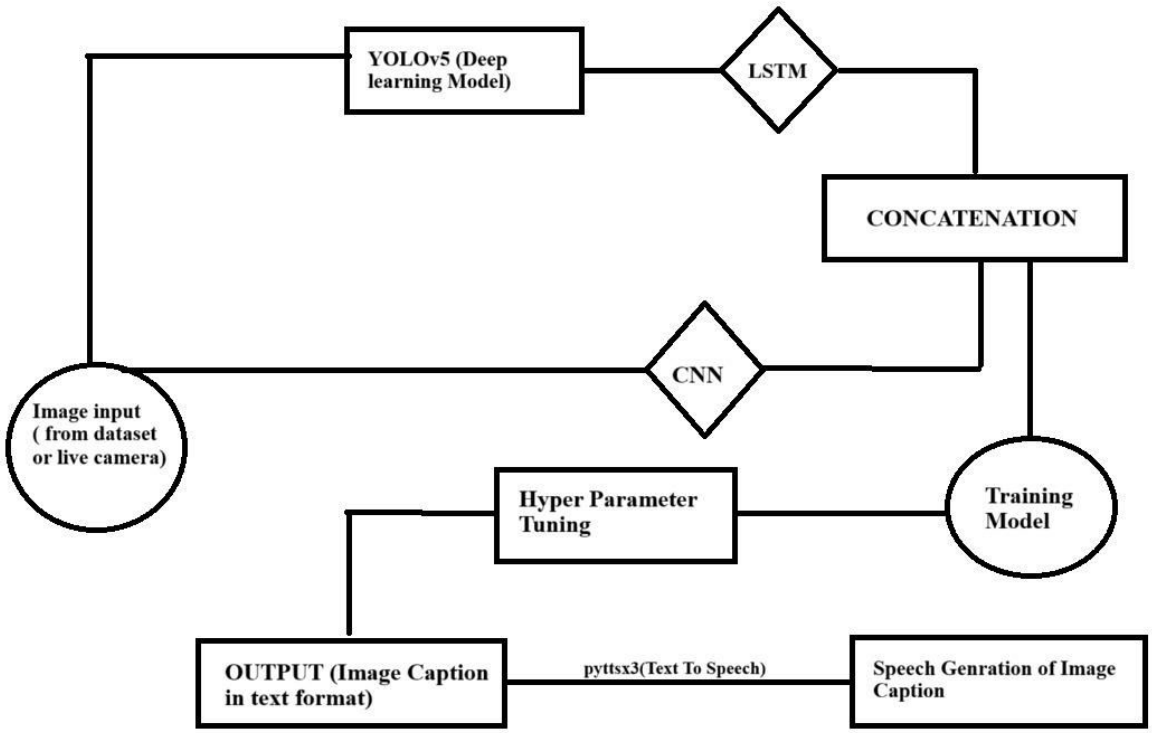


Fig. 1 System Architecture

V. PSEUDO CODE

Step 1: Load the pre-trained models (CNN, Transformer-based model, TTS engine)
Step 2: Gather input data (images from device or webcam)
Step 3: For each image:
   a. Extract visual features using CNN
   b. Generate captions using the Transformer model
   c. Convert captions to speech using TTS
Step 4: Output the audio description to the user
Step 5: Repeat the process for new incoming images

VI. SIMULATION RESULTS

VisionAID system architecture combines several AI technologies to process images and improve user interaction with voice-supported navigation. Images taken or uploaded through a mobile app developed with Flask are saved in a static folder and processed with OpenCV for object detection using the YOLOv5 model. At the same time, the Python Imaging Library (PIL) prepares images for the BLIP model, which produces descriptive captions. The generate_caption() method integrates the output of both models: BLIP generates a general scene description, and YOLOv5 detects particular objects. Spatial reasoning based on bounding box coordinates provides insight into object locations, which allows for an overall arrangement of the scene. Multimodal integration delivers rich, voice-generated descriptions of the environment to visually impaired users.

Along with scene description, the system features a live navigation system that utilizes the Google Maps API to provide turn-by-turn walking instructions. The instructions are read aloud by utilizing Text-to-Speech (TTS) technology. The navigation component also incorporates YOLOv5 to identify proximal obstacles in real-time and provide instant voice warnings to ensure user safety. Voice-controlled from its inception, the system enables users to start, change, or

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305078

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

terminate navigation with easy voice commands. Through the combination of sophisticated path planning and real-time obstacle detection, VisionAID provides visually impaired users with increased independence and safer mobility.

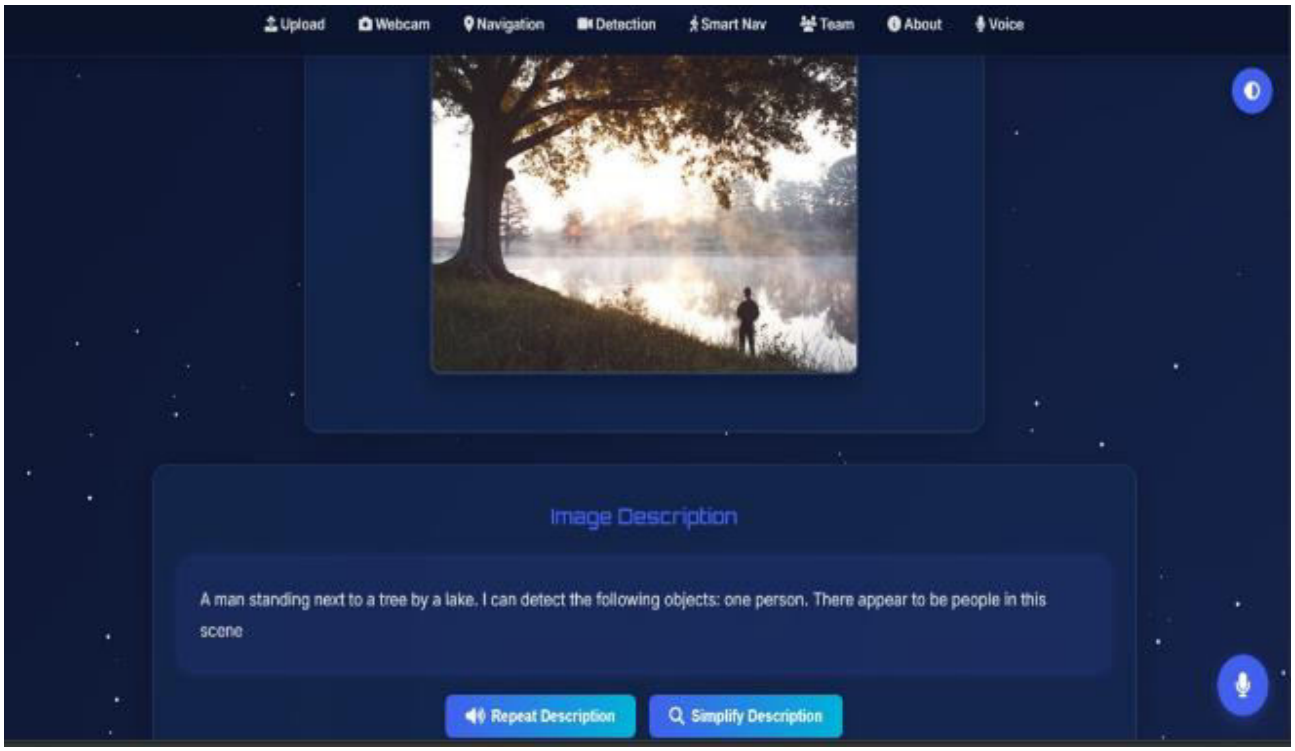


Fig. 2 Generated Image Caption

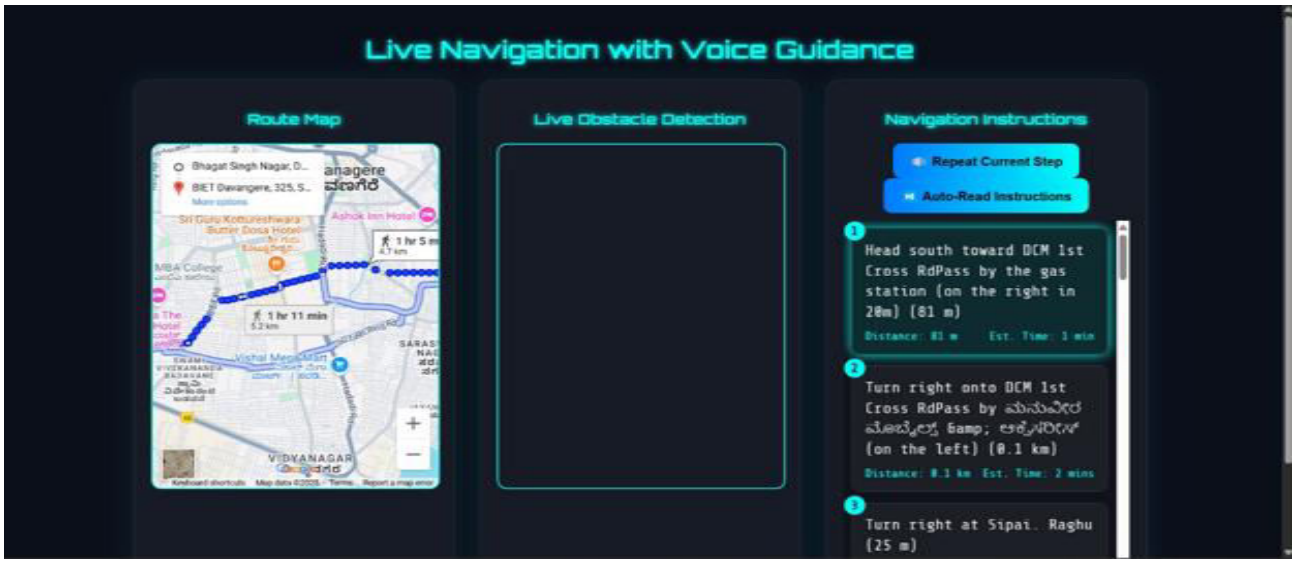


Fig. 3 Live voice-guided navigation with obstacle detection

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305078

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

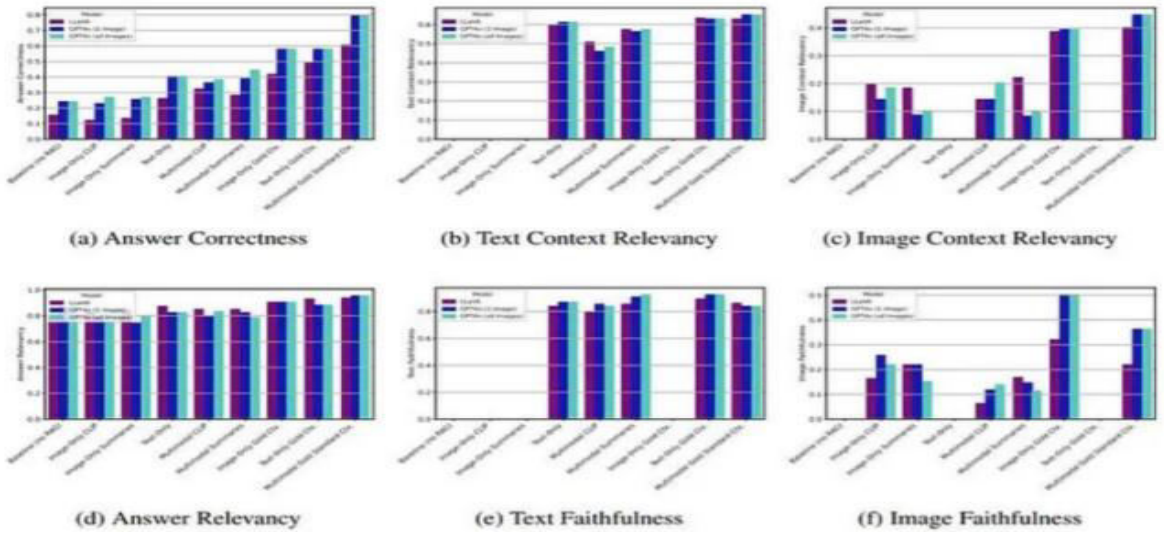


Fig. 4 Evaluation Metrics

VII. CONCLUSION AND FUTURE WORK

VisionAID is a pioneering AI-driven system to facilitate visually blooded individualities with descriptive image captions and real-time voice-activated navigation. Via deep literacy models such as YOLOv5 and BLIP, the system detects objects, scenes, and conditioning from images and transfers visual data to accessible audio description. The system has an easily accessible interface and voice commerce and facilitates availability and autonomy to visually blooded individualities. unborn development entails improved object identification, real- time shadowing, improved voice command interpretation, stoner profile personalization, mobile app integration, and support for native languages. Integration of essential navigation APIs and pall- based processing will also improve delicacy, performance, and usability, transubstantiating VisionAID into a significant and accessible aid for visually impaired individuals.

REFERENCES

1. "Multimodal AI Framework for Visually Impaired Navigation and Description" by Chen and Kumar, Springer DOI: 10.1007/978-981-97-8836-1_22 (2025).
2. "YOLOv5-Based Scene Detection with Voice Alerts for the Blind" by Tanaka and Singh, International Journal of Research Publication and Reviews (IJRPR) (2025).
3. "Transformer-Driven Captioning Systems for Assistive Technology" by Ibrahim and Garcia, Springer, DOI: 10.1007/s11042-024-18966-7 (2024).
4. "AI-Based Mobile App for Real-Time Scene Interpretation" by Mukherjee et al. SSRN (2024).
5. Kommineni, M., & Chundru, S. (2025). Sustainable Data Governance Implementing Energy-Efficient Data Lifecycle Management in Enterprise Systems. In Driving Business Success Through Eco-Friendly Strategies (pp. 397-418). IGI Global Scientific Publishing.
6. "Vision-to-Speech: A Context-Aware Captioning System for Accessibility" by Novak and Sharma, Springer, DOI: 10.1007/s11042-024-20036-x (2023).
7. "Midjourney's Image-to-Text Generator: An Accessibility Tool for the Visually Impaired" by Life wire (2023).
8. "CLIP-Based Vision Assist for Low-Vision Navigation" by Abbas and Zhou, Amazon Science (2023).

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details