International Journal of Technology and Applied Science (IJTAS)
E-ISSN: 2230-9004 ● Website: www.ijtas.com ● Email: editor@ijtas.com

AI-Based Architectural Image Captioning and Voice Generation System for Blind Students

D. Bikshalu[1], P. Shirisha[2], N. Sowmya[3], P. Sai Harshini[4], N. Mounika[5]

[1]Assistant Professor, [2,3,4,5]B. Tech 3rd year Students
[1,2,3,4,5]CSE (AI&ML), Vignan's Institute of Management and Technology for Women, Hyderabad, India.

Abstract:
An AI-powered architectural image captioning and voice generation system is developed to help blind students better understand architectural images by automatically generating textual descriptions. The proposed system employs computer vision and AI algorithms to process images of buildings, rooms, layouts, and architectural structures. Visual features are extracted by Convolutional Neural Networks (CNNs), and Natural Language Processing (NLP) algorithms are used to generate relevant captions. The captions are then transformed into audio by text-to-speech technology. The proposed system enhances visual perception, facilitates inclusive education, and helps blind students better understand architectural concepts independently and effectively.

Keywords: AI, Image Captioning, Computer Vision, Architectural Images, Blind Students, Deep Learning, CNN, NLP, Accessibility, Assistive Technology

I.INTRODUCTION:

Learning in architecture and visual studies is greatly reliant on images, drawings, and visual representation, which pose a challenge to the blind. Architectural concepts like building designs, room designs, elevations, and architectural details are difficult to comprehend without visual exposure. Recent developments in Artificial Intelligence have emerged with promising solutions through image captioning technology. An AI-assisted architectural image captioning system is capable of analyzing architectural images and automatically providing a relevant textual description. By combining computer vision with natural language processing and text-to-speech technology, visual information is transformed into audible content. This is beneficial for independent learning, inclusivity in education, and blind students' better comprehension of architectural concepts and details.

II.RELATED WORK:

Existing studies on image captioning for accessibility have concentrated on the automatic generation of descriptive text from images to assist visually impaired individuals. The early approaches to image captioning were based on the combination of computer vision techniques and rule-based language templates, allowing for the identification of objects and simple descriptions of scenes. However, with the advent of deep learning, encoder-decoder networks based on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) produced more accurate and natural-sounding captions. Examples of such projects include Microsoft's Caption Bot and Google's image description capabilities, which have shown the feasibility of automated image narration, but are general-purpose and lack domain-specific content, such as architectural images.

In the specific domain of architectural visualization, researchers have developed systems for interpreting and classifying architectural images, including the identification of structural elements, floor plans, and design elements. Studies combining semantic image segmentation with building analysis models have enhanced the interpretation of architectural scenes. However, there is a lack of studies specifically

addressing architectural image captioning. Some studies investigate the generation of descriptions for indoor scenes and room layouts, which is relevant to architectural domains, but may not be detailed and technical enough for educational purposes for architecture students.
Recent developments focus on assistive technologies for visually impaired students, integrating image captioning with text-to-speech functionality. Education-specific projects, such as captioning scientific diagrams or math figures, illustrate the need for contextual and domain-specific captioning. Although significant progress has been made, very few projects offer the level of detail necessary for academic learning, with architecture-centric descriptions. This project proposes to leverage these advances by applying deep learning-based captioning models to architectural images, with the goal of providing more accurate and relevant descriptions to benefit blind architecture students.

III. PROPOSED SYSTEM:

The proposed AI-powered architectural image captioning system is intended to help blind students by providing a meaningful text and audio description of architectural images. The system begins by taking an architectural image as input, which can be buildings, floor plans, room designs, or architectural components. A Convolutional Neural Network (CNN) is employed to identify key visual elements such as shapes, objects, and arrangements. The visual elements are then processed by a Natural Language Processing (NLP) model, which can be an LSTM or a Transformer model, to produce a clear and contextually relevant description of the architectural components. The final step involves the text being translated into speech using a text-to-speech module, which helps blind students interpret architectural content independently.

IV. ARCHITECTURE:

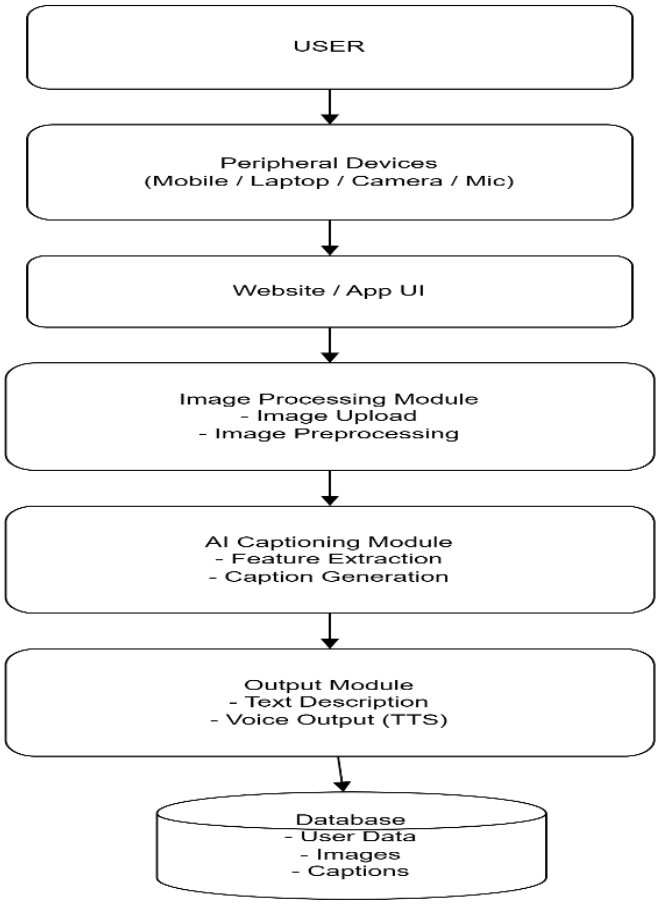

Fig 1: System Architecture

1. Image Input Module
The user, who is a blind student, captures the image using a mobile phone's camera or uploads the captured image. This image is then passed on to the system for processing.
2. Pre-processing Module
The image is resized and normalized. Noise removal and basic enhancement were done. Ensures that the image is valid for feature extraction.
3. Feature Extraction (CNN)
A Convolutional Neural Network (CNN) model like Res Net, VGG extracts important features. Objects, shapes, colors, and other spatial information are identified a feature vector corresponding to the image.
4. Caption Generation
The feature values are passed to the Language Model. LSTM/RNN/Transformer creates a meaningful sentence one word at a time. It combines visual features with language understanding.
5. Text Caption Output
It develops a descriptive caption of the picture.
6. Text-to-Speech
The generated caption is then converted into audio with the help of Text-to-Speech technology. Assists blind students in understanding what they can't see by using sound.
7. Audio Output to User
This is read out by the speaker or headphones. User can listen and understand the image content.

V. IMPLEMENTATION DETAILS:

1. Development Environment
Programming Language: Python
Framework: TensorFlow / Keras or PyTorch
Platform: Desktop / Android using API Integration
IDE: Google Colab / VS Code / PyCharm
2. Dataset Used
- MS COCO/Flickr8k/Flickr30k
- Each image has multiple human-written captions.
- Dataset is divided into:
- Training set
- Validation set
- Testing set

3. Image Pre-processing
- Resized so that they have a certain size, e.g., 224x224.
- Normalization of pixel values.
- Converted into arrays for model input.
- Data augmentation (optional) helps improve accuracy.

4. Feature Extraction using CNN
- Pre-trained CNN models like:
- VGG
- Res Net
- The final classification layer is removed.
- The stored vectors contain the extracted features.
- These vectors contain significant visual information.

5. Text Pre-processing
- Captions are converted to lowercase.
- Special characters and numbers are deleted.

- Tokenization is done.
- Vocabulary is created.
- The sequences are padded to ensure equal lengths.

6. Caption Generation Model

Encoder-Decoder architecture is used:

Encoder: CNN designed for Image Features

Decoder: LSTM / RNN / Transformer

The image features are then sent to the decoder.

The model makes word predictions until an end token is reached.

7. Model Training

Loss Function: Categorical Cross-Entropy

Optimizer: Adam

- Training is performed over multiple epochs.
- Teacher forcing enhances efficiency in learning.

8. Caption Prediction

During testing:

- Image is passed through a CNN.
- In Decoder, caption is generated word by word.
- Greedy search or Beam search is used**.**

9. Text to Speech

- Caption generated will be sent to a TTS engine:
- Google Text-to-Speech
- The audio output is heard by the user.

10. User Interface

- Simple mobile or web interface.
- Buttons for:
- Capture image
- Upload image
- Listen to caption

VI. DEVELOPMENT FRAMEWORK:

1.Programming Language

Used for implementing deep learning models, image processing, and system integration because of its simplicity and rich libraries.

2.Deep Learning Frameworks

Used for developing and training CNN and LSTM/Transformer models for image caption generation.

3. OpenCV

Used for image acquisition, resizing, preprocessing, and camera integration.

4.NLP Framework

Used for text preprocessing, tokenization, and vocabulary generation for caption generation.

5.Text-to-Speech Framework

Used for converting generated text captions into audio output to help blind students

VII. ALGORITHM:

Step 1: Image Captioning

Step 2: Image Preprocessing

Step 3: Feature Extraction

Step 4: Caption Generation

Step 5: Text-to-Speech Conversion
Step 6: Audio Output

VIII. RESULTS:

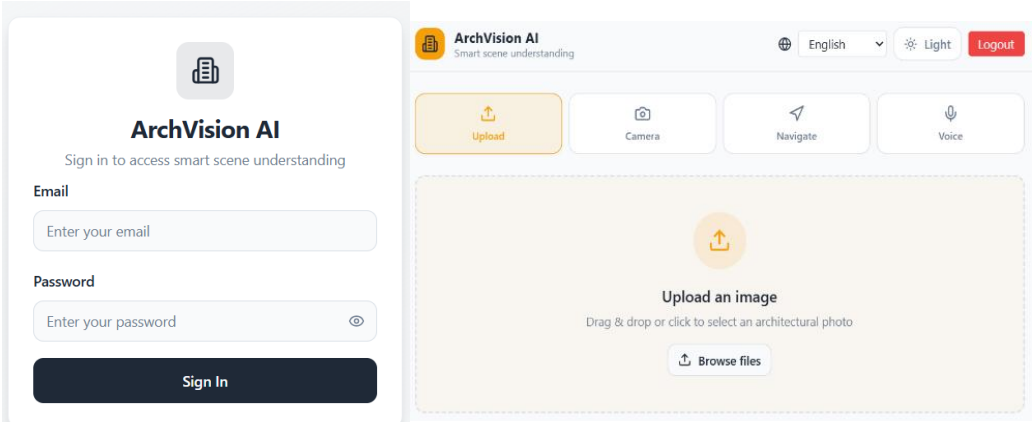

Fig 2: System Interface

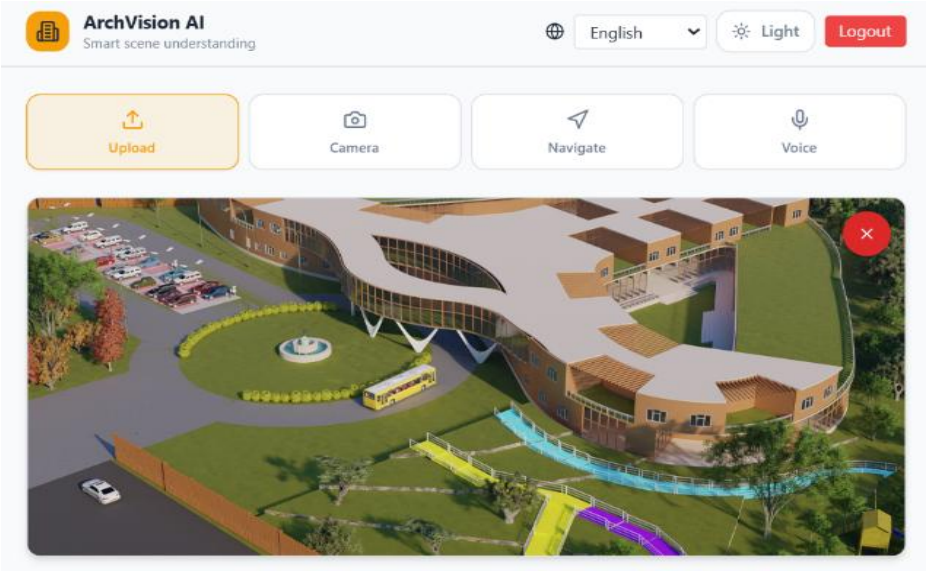

Fig 3: Uploaded Image Display

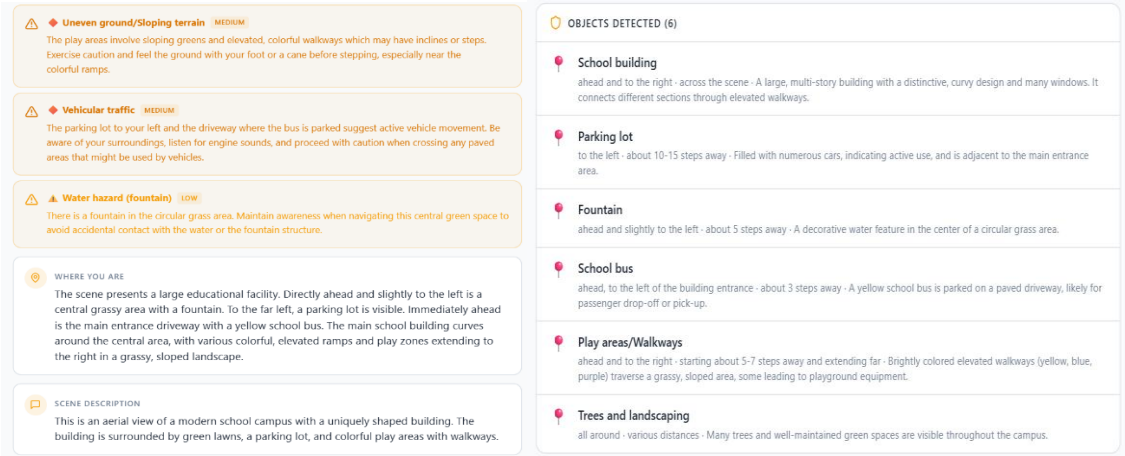

Fig 4: Objects Description

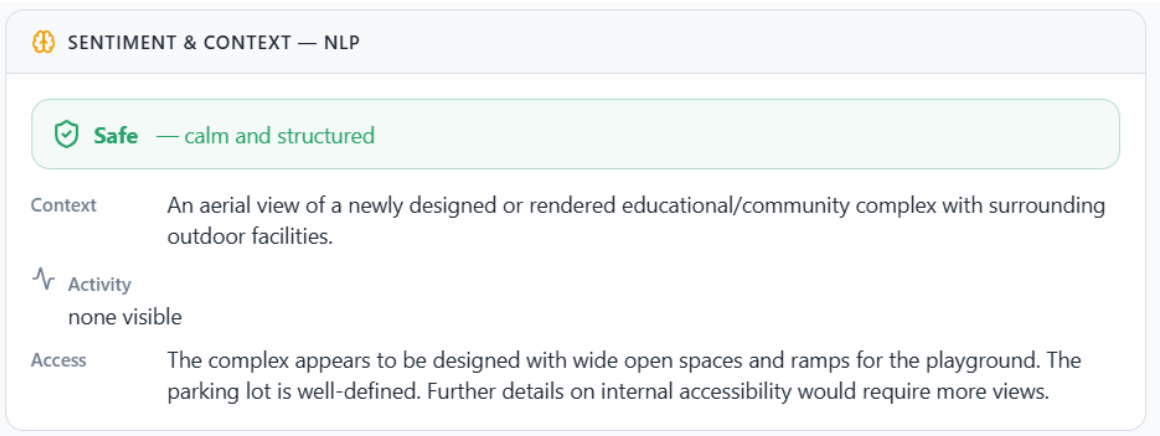

Fig 5: Sentiment & Context NLP

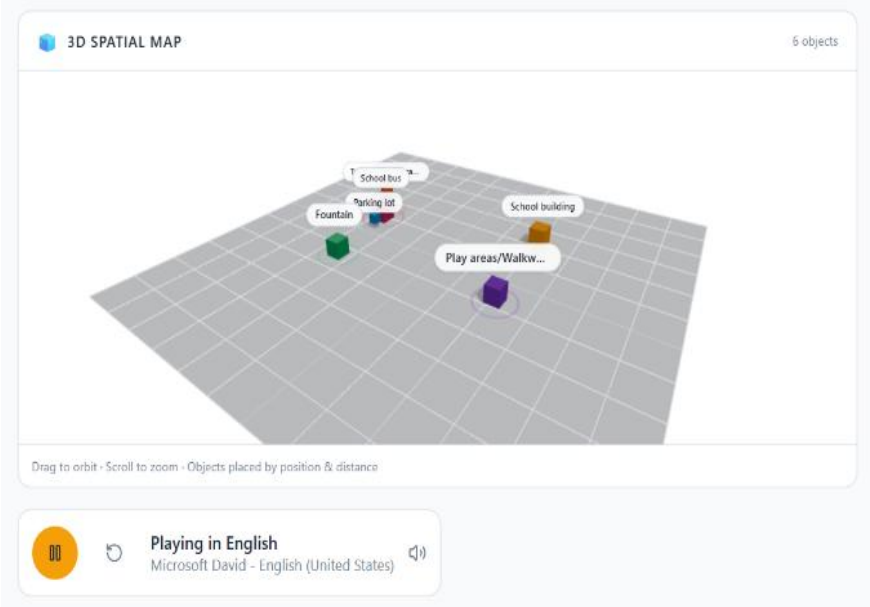

Fig 6: 3D Spatial Map Visualization

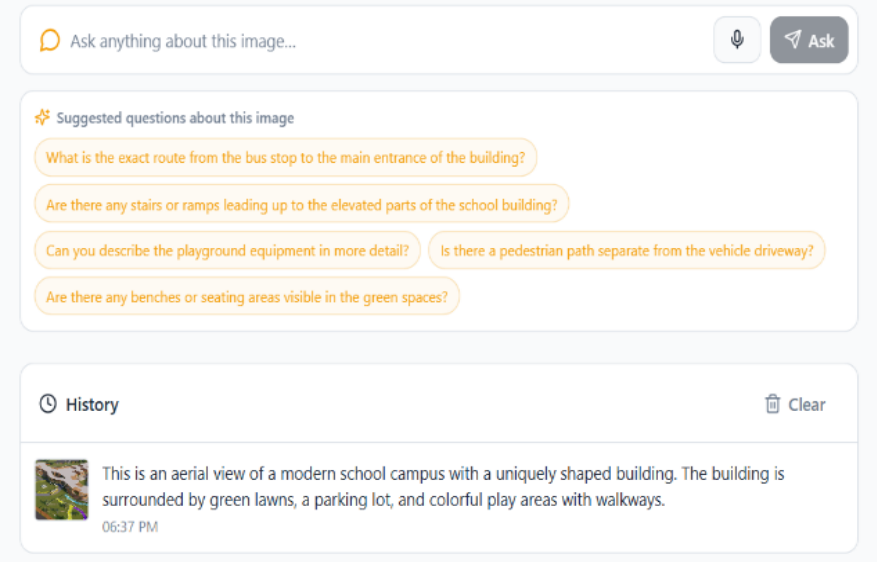

Fig 7: Interactive Captioning and Voice Query Result Screen

X. CONCLUSION:

The AI-powered Image Captioning System effectively closes the gap between visual content and the visually impaired by enabling images to be translated into significant text descriptions and audio output. The system combines computer vision and natural language processing algorithms to effectively recognize objects and scenes in images and produce human-like captions. The text-to-speech functionality of the system allows blind students to access visual information on their own. The system improves accessibility in learning and everyday life. It can be further developed for real-time use and support multiple languages.

REFERENCES:

1. Arthy, S., D. Shanthi, and M. Buvana. "Automatic image annotation and image retrieval in social networks using hadoop." Advances in Natural and Applied Sciences 11, no. 7 (2017): 492-498.
2. Rajasekaran, G., Snowvin, L. D., Vaishnavi, K., Mary, D. A., Rajest, S. S., & Ali, M. M. S. (2025). AI Voice Assistant and Caption Generation Using Convolution Neural Network and Bi LSTM. Central Asian Journal of Mathematical Theory and Computer Sciences, 6(4), 758-77.
3. Venkat Ragavan, S., et al. "A real time portable and accessible aiding system for the blind–a cloud based approach." Multimedia Tools and Applications 82.13 (2023): 20641-20654.
4. Iwamura, Kiyohiko, Jun Younes Louhi Kasahara, Alessandro Moro, Atsushi Yamashita, and Hajime Asama. "Image captioning using motion-CNN with object detection." Sensors 21, no. 4 (2021): 1270.
5. Fernando, Sandra, et al. "Image recognition tools for blind and visually impaired users: An emphasis on the design considerations." ACM Transactions on Accessible Computing 18.1 (2025): 1-21.
6. Wu, H., Yang, H., Chang, F., Zhu, D., & Liu, Z. (2025). AI-generated tactile graphics for visually impaired children: A usability study of a multimodal educational product. International Journal of Human-Computer Studies, 200, 103525.
7. Yadav, M. K. (2025). The role of artificial intelligence in empowering visually impaired students: a comprehensive overview. Natl. J. Res. Innov. Pract, 10, 1-15.
8. Pujala, Vighnesh, and G. Ravi. "AI-Enabled Image Description: Bridging the Gap for the Visually Impaired." (2026).
9. Zahra, M. M. F. Enhancing Accessibility for Visually Impaired Individuals. Diss. 2025.
10. Uikey, Jitesh, et al. "Visual Understanding and Navigation for the Visually Impaired Using Image Captioning." 2025 International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI). IEEE, 2025.
11. Aljedaani, Wajdi, et al. "Visual Aid: Enhancing Accessibility for Visually Impaired Users Through AI." Proceedings of the 22nd International Web for All Conference. 2025.
12. Arun, M. R., et al. "AI Based Smart System to Empower and Assist Persons with Visual Impairment." 2025 2nd International Conference on Electronic Circuits and Signaling Technologies (ICECST). IEEE, 2025.
13. Ainary, Bhanuja. "Audo-Sight: Enabling Ambient Interaction For Blind And Visually Impaired Individuals." arXiv preprint arXiv:2505.00153 (2025).