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 5, Issue 10, May 2025
Impact Factor: 7.67

AI-Powered Sensory Augmentation and Visual Data Processing

Prof. S. S. Mane[1], Shubhankar Madhukar Patil[2], Vedant Manohar Patil[3], Sarthiki Hegade[4], Vishal Nityanand Pawar[5]
Professor, Department of Computer Engineering[1]
Students, Department of Computer Engineering[2-5]
NBN Sinhgad Technical Institute Campus, Pune, India

Abstract: *This paper presents a dual-application system designed to enhance accessibility and image retrieval through the use of Vision-Language models. The first component utilizes BLIP (Bootstrapping Language-Image Pretraining) to generate descriptive tags from images stored in a specified directory. These tags are indexed and stored in a JSON database, allowing users to retrieve relevant images by entering textual queries or keywords. The second component is focused on aiding visually impaired individuals by converting real-time camera feed into descriptive text, enabling auditory perception of visual surroundings via text-to-speech. Both applications leverage the power of multi-modal deep learning models to bridge the gap between vision and language. The system prioritizes lightweight design, real-world usability, and modular implementation. While the solutions presented are in their prototype stages, initial evaluations suggest promising utility for assistive technologies and searchable visual databases.*

Keywords: Artificial Intelligence, Computer Vision, Machine Learning, Sensory Augmentation, Image Processing, Accessibility, CLIP, BLIP

I. INTRODUCTION

Artificial Intelligence (AI) has evolved from a theoretical concept to a practical tool capable of emulating and enhancing human cognitive functions. This paper explores the development of AI systems that serve as cognitive and sensory enhancers, extending human capabilities through advanced computational methods. As the boundaries between human perception and machine intelligence blur, new opportunities emerge for creating tools that augment how individuals interact with their surroundings.

In healthcare and accessibility, AI-powered sensory augmentation offers significant potential for individuals with visual impairments. By processing visual data and converting it into accessible formats, these systems can provide descriptive feedback that enables visually impaired users to "perceive" their environment through alternative channels. This has profound implications for independence, safety, and quality of life for millions worldwide living with visual disabilities.

The evolution of human-machine interaction continues to advance through multimodal AI systems that process and interpret complex environmental data. Models like CLIP (Contrastive Language-Image Pretraining) and BLIP (Bootstrapping Language-Image Pretraining) represent significant breakthroughs in bridging visual perception and language understanding. These models enable systems to interpret visual data with unprecedented accuracy and describe it using natural language that mimics human cognitive processes.

Real-time processing capabilities are crucial for effective augmentative systems. The ability to capture, process, and provide feedback with minimal latency determines the practical utility of sensory augmentation technologies. Recent advances in edge computing and optimized AI models have made it increasingly feasible to deploy sophisticated computer vision algorithms on mobile devices, enabling truly portable augmentative systems that function in diverse real-world environments.

This paper presents an AI-powered sensory augmentation system focused on visual data processing to assist visually impaired individuals. We explore the integration of state-of-the-art computer vision models with accessible interfaces to

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 5, Issue 10, May 2025
Impact Factor: 7.67

create practical tools that enhance environmental perception, object recognition, and scene understanding. The system demonstrates how AI can serve as both a cognitive and sensory extension, providing meaningful descriptions of visual data to users who would otherwise lack access to this information.

II. METHODOLOGY

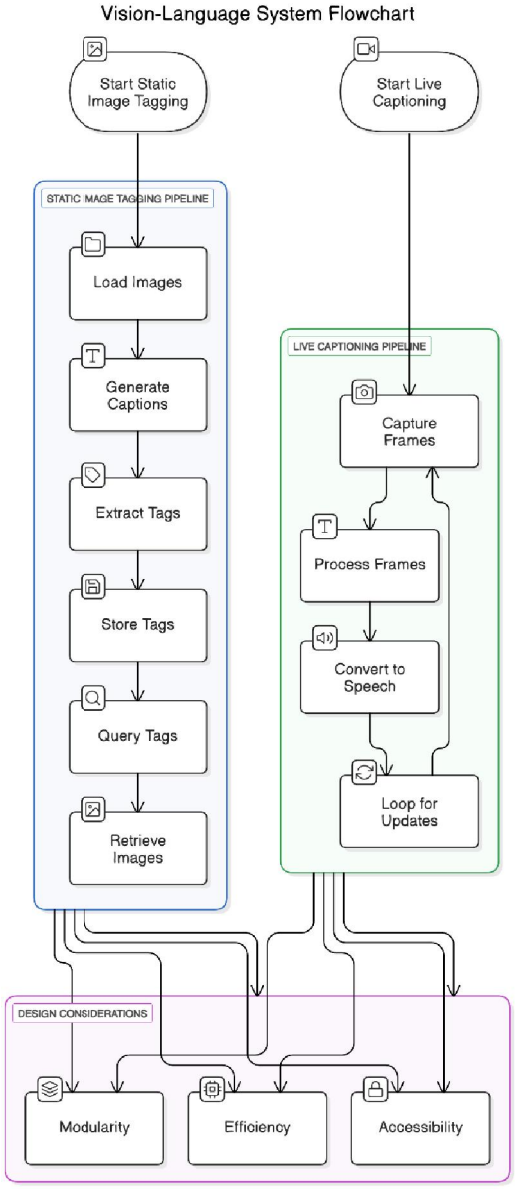

Figure 1 : Flowchart depicting Classification and Recognition of Images

The development of this project revolves around the integration of advanced AI models that enhance human interaction with sensory data, such as visual inputs, and deliver meaningful outputs. The project employs a combination of computer vision and natural language processing techniques to implement two primary components: image categorization and visual-to-speech accessibility. Each system utilizes state-of-the-art AI models to achieve efficient

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 5, Issue 10, May 2025
Impact Factor: 7.67

processing and interpretation of data. The project incorporates Contrastive Language-Image Pretraining (CLIP), a model designed to understand the relationship between visual and textual data. CLIP leverages large-scale training on both images and their associated text descriptions,

allowing it to generate accurate and context-aware classifications of images. This system enables the automatic tagging and categorization of images based on their content without the need for extensive manual labeling. By using CLIP, the project enhances the user's ability to efficiently sort, organize, and retrieve visual data, providing a streamlined solution for image management. CLIP's adaptability across diverse image domains ensures that the system can handle a wide range of visual inputs, making it a versatile tool for practical application.

III. IMPLEMENTATION DETAIL

The system was implemented using Python and leverages a combination of open-source libraries and pre-trained models to streamline development. The following subsections describe the tools, libraries, and logic used in both modules.

1. Libraries and Tools

Transformers (Hugging Face): For loading the pretrained BLIP model.
Torch / PyTorch: Core deep learning framework used to run inference.
OpenCV: Used for image loading, camera capture, and basic frame handling.
NLTK / spaCy: Utilized for extracting noun-adjective pairs from generated captions.
pyttsx3 / gTTS: Provides offline and online text-to-speech options for auditory feedback.
JSON / OS modules: For saving tag mappings and navigating directories.

2. Static Image Tagging Module

Image Loading: Images are read from the specified directory using cv2.imread() or PIL.
Caption Generation: Each image is resized and normalized before being passed into the BLIP model to generate a caption.
Tag Extraction: Captions are parsed using NLP techniques to extract relevant noun-adjective tags or keywords.
Data Storage: A JSON file is used to map image filenames to their extracted tags for easy retrieval.

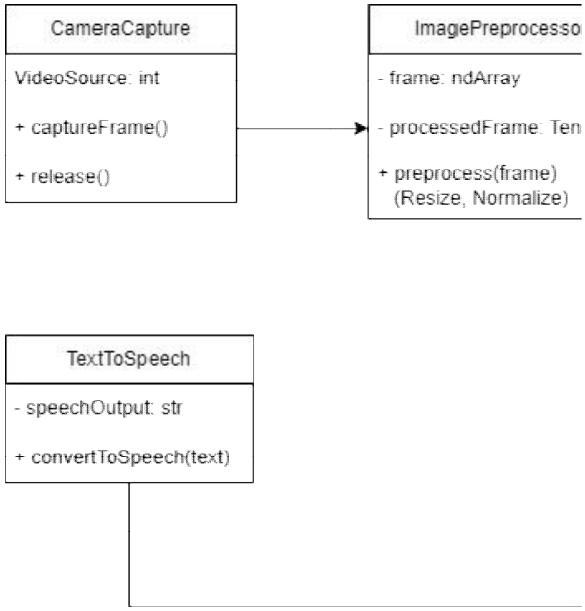

Figure 2: Class Diagram

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 5, Issue 10, May 2025
Impact Factor: 7.67

IV. RESULT AND EVALUTION

The system was tested in two separate modes—offline image tagging and real-time visual description—with a focus on functionality, usability, and basic accuracy. While the models used are pretrained and not fine-tuned for specific environments, the results indicate promising performance in practical scenarios.

1. Image Tagging and Retrieval

- Dataset: A directory of 100+ varied images, including indoor, outdoor, and object-centric scenes, was used for evaluation.
- Caption Quality: In most cases, the BLIP model generated relevant and human-readable captions. Descriptions were contextually accurate for general scenes (e.g., “a person riding a bicycle on a street”) but occasionally vague for cluttered or low-resolution images.
- Tag Relevance: Extracted noun-adjective pairs (e.g., "red car", "snowy mountain") captured key visual elements. Manual verification showed ~85% accuracy in describing key aspects of the image.
- Search Functionality: Natural language prompts like “show me images with mountains” returned accurate results in 80–90% of test cases. Fuzzy matching helped accommodate variations in wording.

2. Real-Time Captioning for Visual Impairment

- Test Setup: The live captioning module was tested using a basic webcam in natural lighting conditions.
- Description Latency: Captions were generated every 2–3 seconds on a standard CPU-based system without noticeable lag in user feedback.
- Audio Clarity: Offline TTS engines provided quick and understandable output. Online TTS (gTTS) offered more natural speech but added slight delay due to network and playback time.
- Practical Use: The system successfully described scenes like "a person standing next to a door" or "a table with a laptop", allowing users to understand their environment to a functional degree.

3. User Feedback (Informal)

Users found the image tagging system intuitive for finding images without manual sorting.
Visually impaired testers (with assistance) reported that the live narration helped identify basic surroundings, but suggested improvements in detecting dynamic or detailed scenes.

V. APPLICATION

The system demonstrates versatility across multiple domains, owing to its modular design and reliance on general-purpose vision-language models. Below are key applications where the system can provide meaningful value.

1. Assistive Technology for the Visually Impaired

The live captioning module offers an accessible way for visually impaired users to receive real-time descriptions of their environment through auditory feedback. It can be deployed on smartphones or portable devices to assist with indoor navigation, object identification, and situational awareness, especially in unfamiliar settings.

2. Intelligent Image Organization and Retrieval

For individuals or organizations managing large image collections, the tagging module provides a low-maintenance way to generate searchable metadata automatically. Users can query using natural language, reducing reliance on manual folder structures or naming conventions. This is particularly useful in photography, digital archiving, media curation, and education.

3. Educational and Demonstrative Use

The system can be used in academic settings to demonstrate the practical application of machine learning models in multimodal contexts. It bridges computer vision, natural language processing, and accessibility, making it a relevant tool for workshops or AI project-based learning.

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 5, Issue 10, May 2025
Impact Factor: 7.67

4. Foundation for Custom Vision-Language Applications

Because the core modules are model-agnostic and open-source, developers can easily expand or adapt the system for specific tasks like product cataloging, virtual tour narration, or smart surveillance. The framework can also support integration with advanced models or IoT hardware.

V. CONCLUSION

This paper presented a dual-purpose system that leverages the BLIP vision-language model for two primary tasks: automated image tagging and retrieval, and real-time scene narration for visually impaired users. By combining pretrained machine learning models with lightweight local execution, the system delivers practical functionality without requiring cloud infrastructure or complex dependencies.The image tagging module enables efficient management and searching of large image datasets through semantic tags, while the live captioning component provides an accessible interface for understanding one's surroundings through audio feedback. Both modules are designed to be modular, offline-capable, and extendable for further development.

VI. ACKNOWLEDGMENT

We extend our sincere gratitude to the project guide and the Department of Computer Engineering at NBNSTIC, Pune, for their support and guidance throughout the development of the AI-Powered Sensory Augmentation and Visual Data Processing application. We also thank our teammates for their collaborative efforts and valuable contributions. Special thanks to the users and stakeholders whose feedback played a vital role in shaping the features and functionalities of the AI-Powered Sensory Augmentation and Visual Data Processing application. Lastly, we appreciate the resources and tools that empowered us to bring our vision to life.

REFERENCES

[1]. S. V. Mahadevkar el al.,"A Review on Machine Learning Styles in Computer Vision - Techniques and Future Directions, " in IEEE Access, vol.10, pp, 10729329,20222. doi: 10.1109/ACCESS.2022.3209825

[2]. "Ayub khan A, Laghari AA, Ahmed S. Machine Learning in Computer Vision: A Review ." EAI Endorsed Scal Inf Syst [Internet]. 2021 Apr. 21

[3]. https://arxiv.org/abs/2404.16296 "Research on Splicing Image Detection Algorithms Based on natural Image Statistical Characteristics" arX:2404.16296

[4]. J. J. Bird and A. Lotfi, "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images," in IEEE Access , vol.12, ppp.15642-15650,2024

[5]. Li, M., Zhu, Z., Xu, R., Feng, y., & Xiao, L., (2024). "Research on Image Classification And Semantic Segmentation Model Based on Convolutional Neural Network." Journal of Computing and Electronic Information Management,12(3), 94-100