Archives available at journals.mriindia.com

International Journal of Recent Advances in Engineering and Technology

ISSN: 2347 - 2812

Volume 14 Issue 02s, 2025

Visiobot: Conversational Image Recognition Chatbot

[1]Anushka Rajemahadik , [2] Aniruddha P. Kshirsagar, [3]Pranita Yewale , [4]Aishwarya Pawar , [5]Mrunali Watkar

[1,3,4,5] Student, Computer Science and Engineering, Karmaveer Bhaurao Patil College Of Engineering, Satara, India.

[2]Assistant Professor, Department of Computer Science and Engineering, Karmaveer Bhaurao Patil College of Engineering Satara, India

Email: anurajemahadik@gmail.com[1], anipk2007@gmail.com[2], pranitayewale2488@gmail.com[3], 992002aishwary@gmail.com[4], mwatkar.2025@gmail.com[5]

Peer Review Information

Submission: 21 Oct 2025

Revision: 18 Nov 2025

Acceptance: 05 Dec 2025

Keywords

Conversational AI, Image Recognition, Chatbot, Natural Language Processing (NLP), Visual Recognition, Real-Time Discussions,, Multimodal Interaction, Deep Learning, Human-Computer Interaction, Real-time image Processing.

Abstract

As artificial intelligence advances, new possibilities for conversational agents have been made possible by the combination of computer vision and natural language processing. A thorough analysis of the VisioBot, a conversational chatbot made for image recognition tasks, is provided in this survey. In order to comprehend how VisioBot processes and interprets images in real-time conversations, we look at the different approaches, frameworks, and architectures used in the system. The survey examines the main obstacles to integrating language and vision models, such as multimodal data processing, captioning images, and visual input user interaction. Furthermore, we evaluate the present state of image recognition chatbots and pinpoint research gaps in areas like scalability, accuracy, and contextual awareness. This survey intends to give researchers and developers working in the domains of artificial intelligence and human-computer interaction a thorough grasp of VisioBot's potential uses and future development by examining recent developments in this area.

Introduction

Artificial intelligence (AI) has created new opportunities for creating more intelligent and interactive systems by combining language and vision. Recent developments in computer vision and natural language processing (NLP) have made it possible for image recognition chatbots to emerge, which can comprehend and react to visual content as well, whereas early chatbots could only process text or voice inputs [1][2

VisioBot is one such sophisticated system; it is a conversational chatbot that uses visual data to interpret images in real time and have meaningful conversations. VisioBot improves the quality of interactions and provides accessibility tools, education, healthcare, and customer service by enabling users to upload or refer to images during conversations. [3] [4].

Combining language and vision, however, comes with a number of difficulties. Among these are managing multimodal data, producing precise captions for images, preserving contextual awareness, and guaranteeing significant answers [5][6]. Many deep learning models have been investigated in recent studies, such as transformer-based models like BERT for language comprehension and convolutional neural networks (CNNs) for image classification [7][9].

In-depth examination of the VisioBot system's

underlying technologies, architectures, and design methodologies is provided in this paper. Additionally, it draws attention to current constraints and research gaps in areas like intelligent response generation, scalability, and accuracy [6][10][11]. A thorough review of conversational image recognition systems as they stand today is intended to stimulate further advancements in AI-powered human-computer interaction.

Literature Review

A chatbot model that employs image recognition to facilitate seamless conversations was presented by P. Swarajya Lakshmi et al. in 2025. Their method combines a chatbot system to answer user inquiries with a convolutional neural network (CNN) for image classification. This facilitates image-based user interaction, increasing the chatbot's accessibility and usability in practical applications [1].

R.R. Kolte et al. (2024) designed a chatbot that can identify objects in an image and generate meaningful responses. The system uses CNNs for image detection and natural language processing (NLP) for generating conversational replies. Their model focuses on improving user experience by allowing image-based interactions, which can be useful in education and assistance services [2]

A chatbot that blends image processing methods with convolutional neural networks was presented by Sailesh

R. et al. (2024). Their system provides descriptive responses after identifying the contents of uploaded images. This enhances usability and engagement by enabling interactive communication in which users can send images rather than text [3].

A chatbot that uses deep learning to process images and produce intelligent responses was presented by Devi Praba N. et al. in 2024. In order to interpret the image and have a conversation, their architecture consists of a dialogue generator and an image recognition module. Visually-driven services like e-commerce and healthcare can benefit from this strategy [4].

Ms. M. Buvana et al. (2024) developed a chatbot that interprets image inputs using AI and deep learning techniques. Their model emphasizes accuracy in identifying image objects and generating human-like replies. It helps in reducing the communication gap between humans and machines in image-based conversations [5].

Sheetal Kusal et al. (2022) conducted a scoping review of AI-based conversational agents, exploring their current technologies and future directions. The paper outlines various chatbot models, including those using vision and language understanding, which support the foundation of image recognition chatbots by explaining their scope, challenges, and potential [6].

Pritham Sriram G. and Prasana Venkatesh S. (2021) introduced an image-classifying chatbot that can interact based on image inputs. They used CNN for image classification and focused on enhancing chatbot interaction through visual content, which helps in building a conversational system that understands pictures [7].

Shengyang Su (2020) shared a final year project that combines image recognition and chatbot technologies. The system classifies objects from images and generates simple responses, demonstrating the feasibility of implementing such models for learning and prototyping purposes [8].

Devlin et al. (2019) proposed BERT, a powerful language model that improved the understanding of natural language. BERT plays a key role in chatbot systems by enhancing the chatbot's ability to comprehend and respond meaningfully to user input, even when integrated with image-based systems [9].

Shafquat Hussain et al. (2019) surveyed chatbot technologies and classification techniques. Their review helps in understanding how different chatbot models are built and the design strategies involved, including conversational agents that can process visual data [10].

Existing System

Many systems have been created that combine image recognition with conversational AI, allowing users to interact using both text and images. These systems help chatbots become smarter and more useful in different situations. Below are some important examples of such systems:

1. Image Classifying AI Chatbot

This system uses image recognition to identify the content of uploaded images. It then gives replies based on the visual data. It is useful for applications where users want to get information just by sending a picture [7].

2. CNN-Based Image Recognition Chatbot

A chatbot was developed using Convolutional Neural Networks (CNNs) to recognize objects in images. After identifying the image, the chatbot answers the user's questions related to it. This method combines image processing with simple natural language processing [3].

3. The Models of BERT and ViLBERT
While ViLBERT is a more sophisticated version that can handle both images and text simultaneously, BERT is a model that aids in understanding the meaning of sentences. Tasks like visual question answering, in which the chatbot examines an image and responds to inquiries about it, are a good fit for these models [9][11].
4. Chatbot with Conversational Image Recognition Users can send an image to this chatbot, and it will ask questions based on it.
5. Evaluation of Conversational Agents Based on AI
A thorough analysis of various chatbots that use both speech and images was conducted. According to the study, chatbots can handle multimodal inputs like text, voice, and images with the help of various tools and designs [6].
6. Visual Recognition by IBM Watson
IBM Watson integrates chatbot technology with image recognition. It can use Watson Assistant to answer user queries and recognize objects in pictures. In professional contexts where comprehending visual data is crucial, it is frequently utilized [4][5].
7. The Visual Dialog System
With this system, users can discuss a single image in real time with a chatbot. The dialogue feels organic and connected because the chatbot retains previous queries and responses. Deep learning is used to comprehend the continuous conversation as well as the images [8].

Proposed System
The suggested system, **VisioBot**, is a conversational chatbot driven by artificial intelligence (AI) that combines natural language processing (NLP) and image recognition to facilitate multimodal human-computer interaction. When a user uploads images during a conversation, VisioBot creates intelligent, context-aware responses based on both textual and visual inputs, in contrast to traditional chatbots that only accept text.

A dual pipeline that simultaneously processes text and image inputs makes up the system architecture. The system employs **CNNs** (Convolutional Neural Networks) for visual comprehension, specifically **YOLO** (You Only Look Once) models for object classification and real-time object detection. It uses transformer-based models for language understanding, like **BERT** or **ViLBERT**, which can process natural language queries and connect them to visual data.
Developed with HTML, CSS, and JavaScript, the frontend presents a responsive and easy chat interface. This covers tools for uploading pictures, running searches, and perusing responses in a conversational sequence. Between the user interface and machine learning models, the backend built in **Python** using **Flask** manages communication.
SpaCy manages natural language chores including parsing, named entity recognition, and tokenization. **OpenCV** is used prior to passing data to deep learning models for image preprocessing chores including resizing and filtering. By means of **TensorFlow** and **PyTorch** frameworks, which enable both training and deployment of NLP and image models, the system's performance is further improved.
One **MySQL** database stores user data including interaction logs and metadata. Employing **Google Cloud** or **AWS**, the system guarantees scalability, high availability, and access to GPU/TPU resources for accelerated model training and inference. **TensorFlow Lite** also lets the system run effectively on edge and mobile devices.
VisioBot uses visual context into dialogues to try to raise the caliber of chatbot interactions. In fields including education, healthcare, accessibility for visually impaired users, and client support, where visual data is so important, it is especially helpful. The system advances AI-driven conversational technologies and solves the difficulty of multimodal understanding.

The system creates a straightforward, precise response a

System Architecture

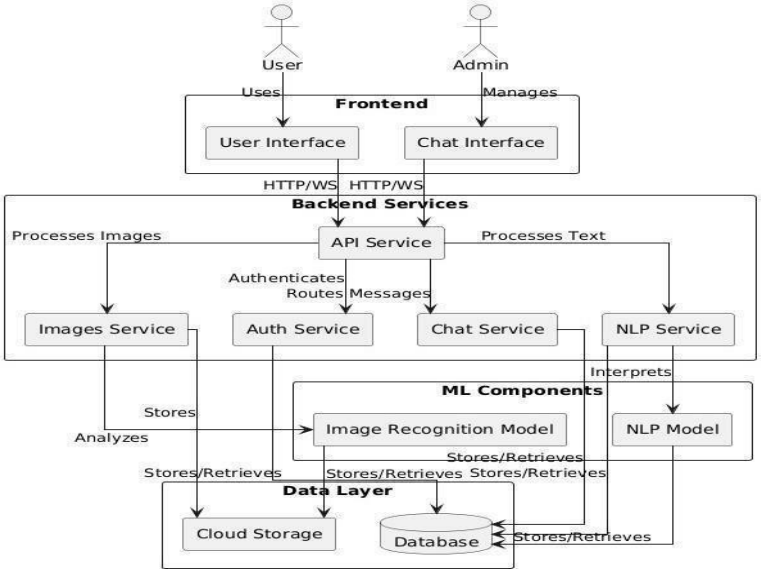

Fig. 1 System Architecture

The system architecture shown in the image represents how different parts of a smart application work together. There are two main users of the system: the User, who interacts with the system through a User Interface, and the Admin, who manages things through a Chat Interface. These interfaces are part of the Frontend, which is what users see and interact with directly.

Behind the scenes, the Backend Services handle all the important tasks. The API Service acts as a bridge between the frontend and backend. The Image Service processes images uploaded by users, while the NLP (Natural Language Processing) Service understands and processes text inputs. The Chat Service manages messages exchanged between the user and the admin, and the Auth Service is responsible for verifying users and securing access

The system also uses Machine Learning (ML) Components to analyze the data. The Image Recognition Model helps understand and analyze visual content, and the NLP Model interprets the meaning of text inputs. These models support the backend services by making the system smarter.

All the information—whether it's images, text, or user details—is stored in the Data Layer, which includes Cloud Storage for large files and a Database for structured information. These storage systems work closely with both the ML models and backend services to store and retrieve data as needed.

Implementation

a. Technology Stack

The development of a conversational image recognition chatbot requires the integration of several technologies that handle both backend machine learning processes and a responsive, user-friendly frontend interface. Below is the comprehensive technology stack used in building the chatbot:

1. Programming Tools

Python: Developing the backend and machine learning models mostly requires Python. Its large ecosystem of libraries and frameworks helps to process images as well as natural language.

JavaScript provides dynamic behaviors and controls communication between the user interface and backend APIs, so driving the interactivity on the frontend.

2. Models of Machine Learning

Deep learning models for image recognition chores as well as natural language processing (NLP) are constructed using TensorFlow. Additionally allowing real-time performance on mobile and edge devices is TensorFlow's interaction with TensorFlow Lite. Renowned for adaptability, PyTorch trains NLP and image processing models, so supporting fast development and testing of deep learning algorithms.

3. NLP, or natural language processing SpaCy: For quick, lightweight NLP tasks like tokenization, parsing, and entity recognition, all essential for deciphering user inquiries.

4. Libraries for Computer Vision OpenCV:

OpenCV handles image processing tasks such as resizing, filtering, and basic feature extraction before images are passed to deep learning models for recognition. YOLO (You Only Look Once): YOLO models are used for real-time object detection, which enables the chatbot to identify and classify objects in images uploaded by the user

5. Data Storage and Management

MySQL: A relational database like MySQL stores structured data, including user information and interaction logs.

6. Cloud and Deployment Platforms

Google Cloud/AWS: Cloud platforms like Google Cloud and AWS are used for training machine learning models, deploying the chatbot, and scaling services. They provide access to compute resources like GPU and TPU instances for faster model training and inference.

Docker: Docker is employed to containerize the chatbot's services, ensuring consistency across different deployment environments and simplifying the scaling process.

7. Frontend Design with HTML and CSS

The front-end of the chatbot is crucial for providing an engaging and intuitive user experience. HTML structures the user interface, while CSS is used to style the layout and ensure responsiveness across devices. Together, they create an interface where users can interact with the chatbot in real time, upload images, and receive responses seamlessly

Key elements of the front-end design include

- Chat Interface: Designed with HTML and CSS, the chat window consists of a message area where user queries and bot responses are displayed. It also includes an input box for users to type queries and upload images.
- Responsive Layout: CSS ensures that the chatbot interface adapts to different screen sizes, providing a consistent experience across desktops, tablets, and mobile devices.
- Message Bubbles: CSS is used to style the message bubbles, with distinct designs for user messages and chatbot responses, improving the visual experience.

Interactive Input: The chat input field and buttons are styled to encourage user interaction, with buttons for sending messages or uploading images

8. APIs and Integration

Flask: These lightweight web frameworks enable the backend to handle HTTP requests from the front-end interface, process images and text, and return responses in real time.

9. Real-Time Processing and Websockets

Socket.IO: Used for real-time communication between the user and the chatbot, ensuring that messages are exchanged instantly, mimicking real-time conversation behavior.

b. Frontend Web Interface and UI Screenshots

The frontend of the chatbot is built to provide an intuitive, user-friendly experience, allowing users to interact seamlessly with the system through text and image uploads.

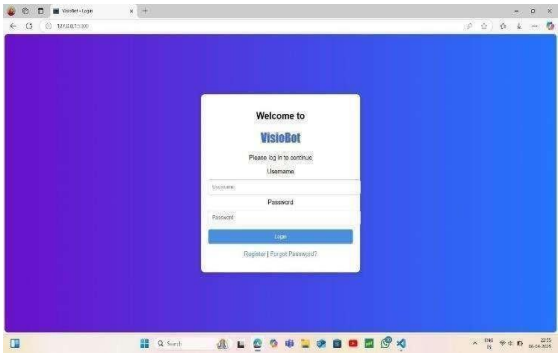

Fig.1.1. Login

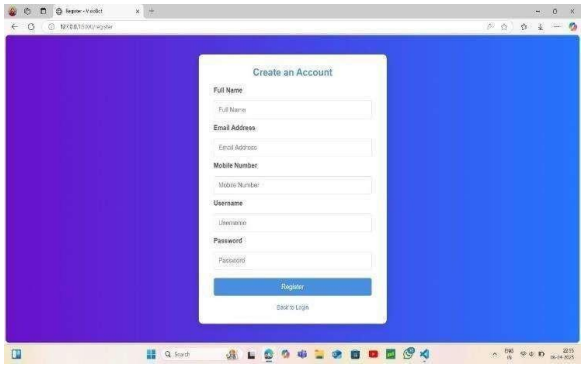

Fig.1.2 User registration form to create a new account

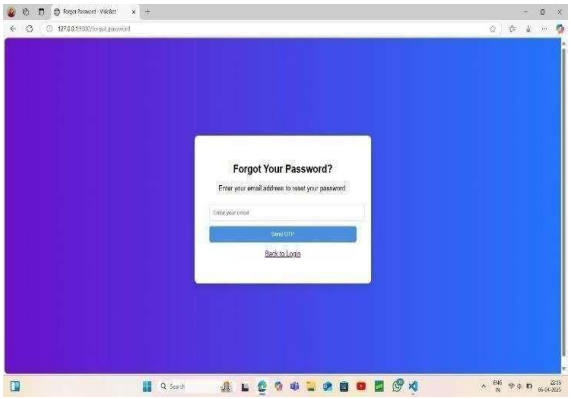

Fig.1.3 Form to enter email to reset password via OTP

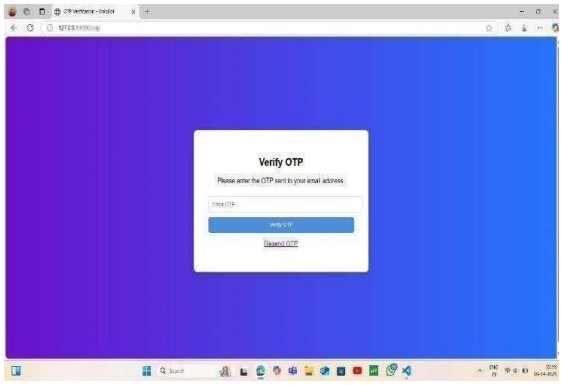

Fig.1.4 Page to enter OTP received in email

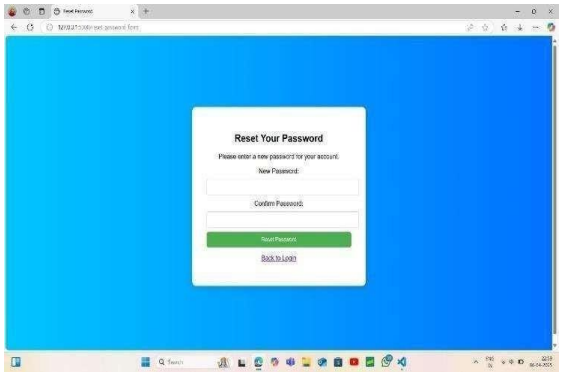

Fig.1.5 Page to enter and confirm new password

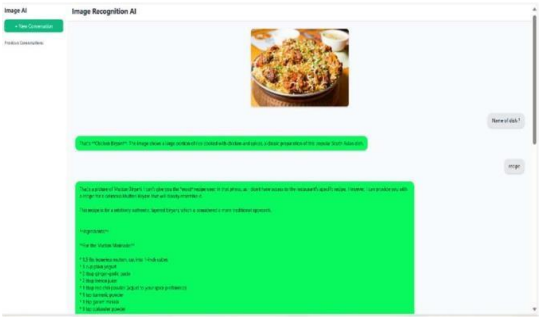

Fig. 1.6 Chat Interface

Conclusion

The exploration of conversational AI combined with image recognition technologies highlights the significant strides made in enhancing human- computer interaction. The integration of natural language processing with computer vision enables chatbots to handle both visual and textual inputs, leading to more intuitive and effective user interactions. While current systems demonstrate impressive capabilities, challenges such as real-time performance, data fusion, and ethical considerations still need to be addressed. Continued research and development in multimodal AI and real-time optimization will be critical in advancing these systems. As they evolve, conversational image recognition chatbots are expected to play transformative roles in various sectors, including healthcare, education, and customer service.

References

P.Swarajya Lakshmi, Shaistha Muskan, Mathangi Anusha, Akireddy Anusha "Conversational Image Recognition Chatbot" March 2025.

Prof. R.R. Kolte, Harsh Wanwe, Prajwal Sathawane, Sahil Kumbhare, Rohit Nagrikar "CONVERSATIONAL IMAGE RECOGNITION CHATBOT" https://www.irjmets.com/uploadedfiles/paper/ /issue_10_october_2024/62967/final/fin_irjmets1 7301 75122.pdf October-2024

Sailesh R , Subiksha S, Yamini R , Mr.Dhinakaran, Naveenkumar K "Convolutional and Image Recognition Chatbot" November 2024

Devi Praba N , Kishor Kumar S , Vijay Surya R , Ananya R Melagiri4 , Vijayaraman S "Conversational Image Recognition Chatbot" https://ijatem.com/wp-content/uploads/2024/11/ICRCCT24_001.pdf November 2024

Ms. M. Buvana, P. Krishna Bharathi, H. Ammar, S. Balaji "Conversational Image Recognition Chatbot" https://www.ijeijournal.com/papers/Vol13-Issue11/1311133138.pdf November 2024

SHEETAL KUSAL,SHRUTI PATIL,JYOTI CHOUDRIE,KETAN KOTECHA,SASHIKALA MISHRA,AJITH ABRAHAM "AI-Based Conversational Agents:A Scoping Review from Technologies to Future Directions" January 2022

Pritham Sriram G., Prasana Venkatesh S. "Image Classifying AI Chatbot" 2021

Shengyang Su "Final Project Report Conversational Image Recognition Chatbot"2020

Devlin, J. et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" 2019

Shafquat Hussain,Omid Ameri Sianaki,and Nedal Ababneh "A Survey on Conversational Agents/Chatbots Classification and Design Techniques" 2019

Lu, J. et al. "ViLBERT: Pretraining Task- Agnostic Visiolinguistic Representations for Vision- and-Language Tasks" 2019