Design and Development of an AI-Powered Multimodal Humanoid Robot (AI Bot) with Integrated Audio-Visual Intelligence and Internet-Based Knowledge Retrieval

Mr. Digeshwar Prabhakar Rajgade

Student, Computer Science and Egineering, Student

Abstract

This paper presents the design and development of a fully equipped humanoid robot named AI Bot. It supports voice interaction, vision, mechanical movement, and internet-based intelligence. The platform combines a Raspberry Pi 5 for AI processing and online APIs, ESP32-CAM for visual sensing, and Arduino Mega 2560 + RAMPS 1.4 for motor control. The AI Bot can answer verbal queries, perform physical gestures, and access live content from Google Search, YouTube, and Wikipedia to respond intelligently. It is powered by a 12 V, 400 W, 400 A supply, ensuring sufficient current for servo and stepper operations. The robot can access the internet via a GSM module, enabling wireless connectivity in mobile scenarios. This makes AI Bot suitable for use in education, assistance, and interactive robotics.

Keywords: Raspberry Pi 5, ESP32-CAM, Arduino Mega 2560, RAMPS 1.4, Stepper Motor, Servo Moto, Google Search API, YouTube Control, Wikipedia Summarization, GSM Internet Access, Humanoid Robot

1. Introduction

As voice assistants and AI continue evolving, combining them with physical robots creates rich, interactive experiences. Unlike purely software-based agents (e.g., Alexa, Siri), the AI Bot uses real hardware to sense, speak, move, and understand. Key capabilities include: Voice-to-Action AI using Python on Raspberry Pi Vision input using ESP32-CAM for object/face tracking Physical motion using stepper and servo motors Live internet knowledge access using Google Search, Wikipedia summarization, and YouTube API Mobile connectivity using GSM internet access This hybrid embodiment enables both emotional and informational engagement with users.

2. Literature Review

Virtual Assistants (Google Assistant, Alexa) respond vocally but lack a body or camera vision. Open-source bots like Mycroft and Jarvis offer NLP but miss physical embodiment. ESP32-CAM boards are often used for streaming vision data to a host. For controling the Stepper motors used Arduino Mega 2560 connected with RAMPS 1.4 shild Integration of Google Search, Wikipedia, and YouTube APIs in physical robots is still a novel approach and rarely seen in academic or hobby projects. GSM-based IoT solutions are popular in mobile robots but rarely integrated with full-featured AI bots.

3. Methodology

3.1 Speech Recognition

- Tool: Google Speech API / Whisper
- Method: Convert voice to text
- Subsystem Identification
- Voice input/output via mic & speaker

3.2 Natural Language Understanding

- Tool: OpenAI GPT / Rasa
- Function: Understand intent, decide action

3.3 Web Integration

- Google Search via SerpAPI
- YouTube Data API
- Wikipedia Python Library

3.4 Computer Vision

- Tool: OpenCV / YOLO
- Use: Face detection, object tracking

3.5 Text-to-Speech

- Tool: pyttsx3 / gTTS

3.6 Hardware Control

- Motor: Arduino or Raspberry Pi
- Interface: pyserial, GPIO
- Vision via ESP32-CAM
- Actuation via Arduino Mega 2560 + RAMPS 1.4
- Internet access through APIs and GSM
- Component Responsibilities
- Arduino: Servo and stepper control
- ESP32-CAM: Camera input over Wi-Fi

4. System Architecture

4.1 Hardware Architecture

- Raspberry Pi 5: Controls speech recognition, APIs, decision logic
- ESP32-CAM: Real-time face/object tracking via Wi-Fi stream
- Arduino Mega 2560 + RAMPS 1.4: Drives stepper and servo motors
- SG90/MG996R servos: Eyes, lips, hands
- NEMA17 stepper motors: Head, arm, or leg movement
- Mic and Speaker: Voice input/output
- Power Supply: 12 V, 400 W, 400 A for motors; 5 V regulator for logic
- GSM Module: Enables wireless internet access for Google, YouTube, and Wikipedia integration
- Communication:
- USB Serial (Raspberry Pi ↔ Arduino)
- HTTP stream (ESP32-CAM ↔ Pi)
- Internet APIs (Google, Wikipedia, YouTube)

5. Working Flow

5.1 Programming Flow

Internet Features

Google Search using googlesearch-python

Wikipedia summaries using wikipedia Python lib

YouTube content titles fetched via youtube-search-python and narrated

Integration and Testing

Commands tested with various accents and languages

Servo gestures matched with verbal feedback

The following diagram represents the structured step-by-step workflow of the AI Bot, integrating Google, YouTube, and Wikipedia functionalities:

5.2 Figure: Box-Structured Workflow Diagram

6. Implementation

6.1 Python Libraries:

- speech_recognition
- pyttsx3
- serial
- OpenCV
- Wikipedia
- youtube-search-python
- googlesearch-python

6.2 Arduino Code:

- C++ based
- using Servo.h
- AccelStepper.h

6.3 ESP32-CAM:

- Arduino IDE
- using face/object detection
- firmware

6.4 Power:

- Single 12 V 400 W SMPS with step-down modules for 5 V logic

6.5 Speech Output:

- Either pyttsx3 (offline)
- gTTS (online) for natural voice

6.6Internet:

- Through Ethernet-Wi-Fi
- GSM Module

7. Results

- Speech recognition: ~94% accuracy
- Real-time search success (Google/Wiki/YouTube): 100%

- Motion sync with speech: 92% accurate
- Camera detection: ~2.5 meters face tracking
- GSM-based access tested on mobile hotspot networks

8. Comparison: Physical Bot vs Virtual Bot

FeatureAI Bot (Physical) Virtual Assistant

Voice Interaction	✅ Yes	✅ Yes
Physical Movement	✅ Yes (Motors)	❌ No
Real-time Vision	✅ ESP32-CAM	❌ No
Google/Wiki/YouTube	✅ Integrated	✅ (only text/audio)
Embodiment	✅ Humanoid	❌ Virtual Only
Emotional Presence	✅ Gestures + Voice	⚠️ Limited to voice
Educational Use	✅ Demonstrable	⚠️ Voice-only
Answering Capability	⚠️ Limited by memory (can't answer all questions)	✅ Can answer most queries using internet

9. Applications

- Education – Interactive teaching assistant using Wikipedia/YouTube
- Elderly Care – Voice-controlled, emotionally expressive helper
- STEM Demonstrator – Teaching motion control and search together
- Event Host/Guide – Gesture and voice-driven live interaction

10. Conclusion

The AI Bot represents a convergence of physical robotics, multimodal interaction, and cloud intelligence. By integrating Google, YouTube, and Wikipedia APIs with a humanoid interface and GSM-based connectivity, it goes beyond traditional bots. With accurate sensing, voice capability, and emotional gestures, it can serve as a powerful tool in both learning and assistive domains.

11. Future Scope

- Add offline LLM (e.g., LLaMA or Whisper) for edge AI
- Enable gesture recognition using ESP32-CAM stream
- Support real-time news & multilingual Wikipedia
- Add emotional TTS and more natural dialogue