JLISNT Journal of Leading Innovations in Science and Technology

Journal of Leading Innovations in Science and Technology, Vol.1, No.3, pp:112-131, July 2026.

EVA: An Intelligent Voice-Controlled Robotic Assistant Integrating ESP32-CAM-Based Vision and Spatial Memory

Vaidik Tiwari[1]*, Mr. Jagdish Pampania[2]

[1 & 2]Department of Mechatronics Engineering, Parul University, Gujrat, India.

Email: onduty670@gmail.com, jagdishpampania2010@gmail.com

***Abstract*-This paper presents EVA, an intelligent voice-controlled robotic assistant integrating speech recognition, ESP32-CAM vision, spatial memory and a multi-degree-of-freedom robotic arm. The system combines perception, decision making and actuation through a modular architecture. Continuous listening, interrupt-driven control and asynchronous software enable responsive human–robot interaction. The robotic arm executes basic manipulation tasks while the vision module detects objects and supports contextual awareness. The platform demonstrates the feasibility of integrating embedded AI, computer vision and robotics into a unified assistant suitable for laboratories, smart homes and assistive environments.**

Keywords-Voice assistant, ESP32-CAM, robotic arm, embedded AI, computer vision, spatial memory, Arduino, human–robot interaction.

I. Introduction

Artificial intelligence and embedded robotics have enabled the development of interactive robotic assistants capable of performing both digital and physical tasks. Conventional voice assistants are largely restricted to software operations, whereas robotic manipulators generally execute predefined motions without contextual awareness. EVA integrates these domains into a unified architecture by combining voice interaction, computer vision and robotic manipulation. The system continuously monitors user commands, processes environmental information and executes appropriate responses. The design emphasizes modularity, scalability and real-time operation. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that

Corresponding Author: Vaidik Tiwari
Email id: *onduty670@gmail.com*

Received: Feb. 2026; Accepted: June 2026; Published online: July 2026

combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

Artificial intelligence and embedded robotics have enabled the development of interactive robotic assistants capable of performing both digital and physical tasks. Conventional voice assistants are largely restricted to software operations, whereas robotic manipulators generally execute predefined motions without contextual awareness. EVA integrates these domains into a unified architecture by combining voice interaction, computer vision and robotic manipulation. The system continuously monitors user commands, processes environmental information and executes appropriate responses. The design emphasizes modularity, scalability and real-time operation. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

Artificial intelligence and embedded robotics have enabled the development of interactive robotic assistants capable of performing both digital and physical tasks. Conventional voice assistants are largely restricted to software operations, whereas robotic manipulators generally execute predefined motions without contextual awareness. EVA integrates these domains into a unified architecture by combining voice interaction, computer vision and robotic manipulation. The system continuously monitors user commands, processes environmental information and executes appropriate responses. The design emphasizes modularity, scalability and real-time operation. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

Artificial intelligence and embedded robotics have enabled the development of interactive robotic assistants capable of performing both digital and physical tasks. Conventional voice assistants are largely restricted to software operations, whereas robotic manipulators generally execute predefined motions without contextual awareness. EVA integrates these

Corresponding Author: Vaidik Tiwari
Email id: onduty670@gmail.com

Received: Feb. 2026; Accepted: June 2026; Published online: July 2026

domains into a unified architecture by combining voice interaction, computer vision and robotic manipulation. The system continuously monitors user commands, processes environmental information and executes appropriate responses. The design emphasizes modularity, scalability and real-time operation. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

Artificial intelligence and embedded robotics have enabled the development of interactive robotic assistants capable of performing both digital and physical tasks. Conventional voice assistants are largely restricted to software operations, whereas robotic manipulators generally execute predefined motions without contextual awareness. EVA integrates these domains into a unified architecture by combining voice interaction, computer vision and robotic manipulation. The system continuously monitors user commands, processes environmental information and executes appropriate responses. The design emphasizes modularity, scalability and real-time operation. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

Artificial intelligence and embedded robotics have enabled the development of interactive robotic assistants capable of performing both digital and physical tasks. Conventional voice assistants are largely restricted to software operations, whereas robotic manipulators generally execute predefined motions without contextual awareness. EVA integrates these domains into a unified architecture by combining voice interaction, computer vision and robotic manipulation. The system continuously monitors user commands, processes environmental information and executes appropriate responses. The design emphasizes modularity, scalability and real-time operation. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that

Corresponding Author: Vaidik Tiwari
Email id: onduty670@gmail.com
Received: Feb. 2026; Accepted: June 2026; Published online: July 2026

combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

Artificial intelligence and embedded robotics have enabled the development of interactive robotic assistants capable of performing both digital and physical tasks. Conventional voice assistants are largely restricted to software operations, whereas robotic manipulators generally execute predefined motions without contextual awareness. EVA integrates these domains into a unified architecture by combining voice interaction, computer vision and robotic manipulation. The system continuously monitors user commands, processes environmental information and executes appropriate responses. The design emphasizes modularity, scalability and real-time operation. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

Artificial intelligence and embedded robotics have enabled the development of interactive robotic assistants capable of performing both digital and physical tasks. Conventional voice assistants are largely restricted to software operations, whereas robotic manipulators generally execute predefined motions without contextual awareness. EVA integrates these domains into a unified architecture by combining voice interaction, computer vision and robotic manipulation. The system continuously monitors user commands, processes environmental information and executes appropriate responses. The design emphasizes modularity, scalability and real-time operation. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

Corresponding Author: Vaidik Tiwari
Email id: ondutyб70@gmail.com

Received: Feb. 2026; Accepted: June 2026; Published online: July 2026

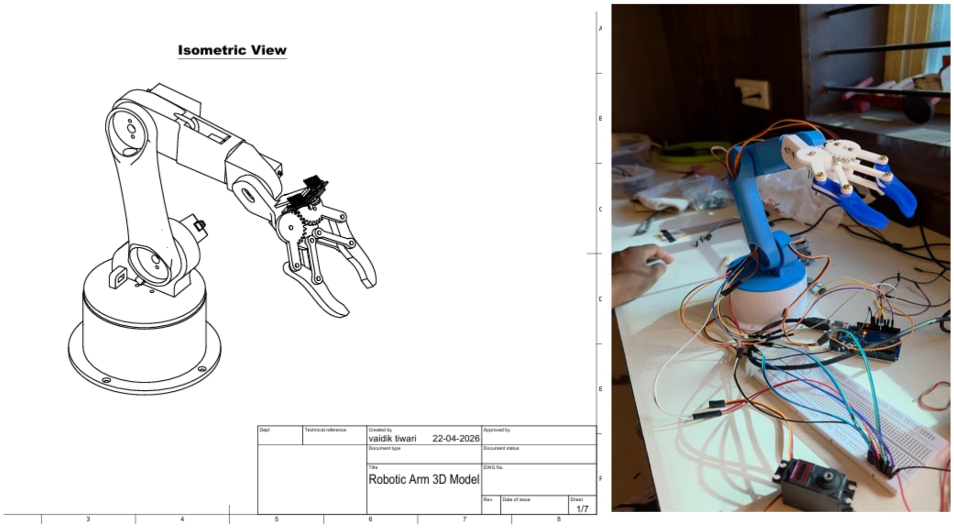

Figure 1. Prototype of the developed robotic arm.

II. System Architecture

The proposed platform consists of a robotic arm, Arduino controller, ESP32-CAM, microphone, speaker and power subsystem. The architecture follows perception–processing–action principles. Voice commands are processed by the software engine, visual information is obtained through the ESP32-CAM, and motion commands are transmitted to servo motors. Separate software modules handle speech recognition, synthesis, interrupt management and robotic control. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

The proposed platform consists of a robotic arm, Arduino controller, ESP32-CAM, microphone, speaker and power subsystem. The architecture follows perception–processing–action principles. Voice commands are processed by the software engine, visual information is obtained through the ESP32-CAM, and motion commands are transmitted to servo motors. Separate software modules handle speech recognition, synthesis, interrupt management and robotic control. The proposed methodology

Corresponding Author: Vaidik Tiwari
Email id: ondutv670@gmail.com

Received: Feb. 2026; Accepted: June 2026; Published online: July 2026

emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

The proposed platform consists of a robotic arm, Arduino controller, ESP32-CAM, microphone, speaker and power subsystem. The architecture follows perception–processing–action principles. Voice commands are processed by the software engine, visual information is obtained through the ESP32-CAM, and motion commands are transmitted to servo motors. Separate software modules handle speech recognition, synthesis, interrupt management and robotic control. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

The proposed platform consists of a robotic arm, Arduino controller, ESP32-CAM, microphone, speaker and power subsystem. The architecture follows perception–processing–action principles. Voice commands are processed by the software engine, visual information is obtained through the ESP32-CAM, and motion commands are transmitted to servo motors. Separate software modules handle speech recognition, synthesis, interrupt management and robotic control. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

The proposed platform consists of a robotic arm, Arduino controller, ESP32-CAM, microphone, speaker and power subsystem. The architecture follows perception–processing–action principles. Voice commands are processed by the software engine, visual information is obtained through the ESP32-CAM, and motion commands are transmitted to servo motors. Separate software modules handle speech recognition, synthesis, interrupt management and robotic control. The proposed methodology

Corresponding Author: Vaidik Tiwari
Email id: ondutty670@gmail.com
Received: Feb. 2026; Accepted: June 2026; Published online: July 2026

emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

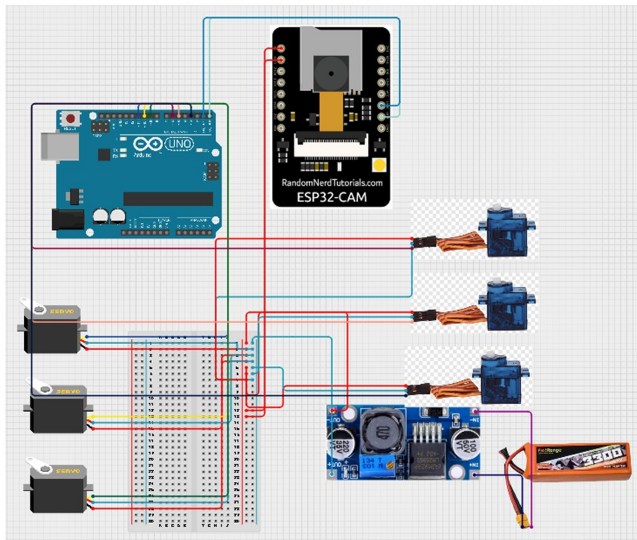

Figure 2. Electronic architecture of the proposed system.

The proposed platform consists of a robotic arm, Arduino controller, ESP32-CAM, microphone, speaker and power subsystem. The architecture follows perception–processing–action principles. Voice commands are processed by the software engine, visual information is obtained through the ESP32-CAM, and motion commands are transmitted to servo motors. Separate software modules handle speech recognition, synthesis, interrupt management and robotic control. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

The proposed platform consists of a robotic arm, Arduino controller, ESP32-CAM, microphone, speaker and power subsystem. The architecture follows perception–processing–action principles. Voice commands are processed by the software engine, visual information is obtained through the ESP32-CAM, and motion commands are transmitted to servo motors. Separate software modules handle speech recognition,

Corresponding Author: Vaidik Tiwari
Email id: ondutv670@gmail.com
Received: Feb. 2026; Accepted: June 2026; Published online: July 2026

synthesis, interrupt management and robotic control. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

The proposed platform consists of a robotic arm, Arduino controller, ESP32-CAM, microphone, speaker and power subsystem. The architecture follows perception–processing–action principles. Voice commands are processed by the software engine, visual information is obtained through the ESP32-CAM, and motion commands are transmitted to servo motors. Separate software modules handle speech recognition, synthesis, interrupt management and robotic control. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

III. Methodology

The robotic arm employs multiple servo motors providing several degrees of freedom for positioning and grasping. A lightweight end-effector carries the ESP32-CAM, enabling eye-in-hand perception. An external battery with a voltage regulator powers the actuators while maintaining stable controller operation. PWM signals control the servo positions, allowing repeatable movement. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

The robotic arm employs multiple servo motors providing several degrees of freedom for positioning and grasping. A lightweight end-effector carries the ESP32-CAM, enabling eye-in-hand perception. An external battery with a voltage regulator powers the actuators while maintaining stable controller operation. PWM signals control the servo positions,

Corresponding Author: Vaidik Tiwari
Email id: ondutv670@gmail.com
Received: Feb. 2026; Accepted: June 2026; Published online: July 2026

allowing repeatable movement. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

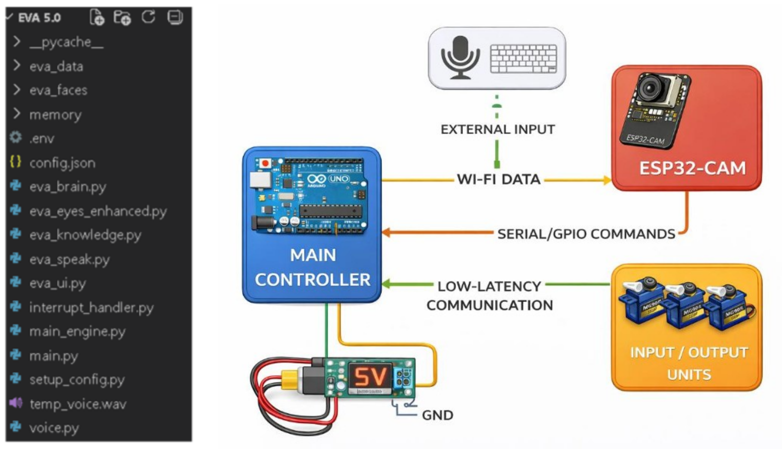

Figure 3. Software and Communication architecture with asynchronous modules.

The robotic arm employs multiple servo motors providing several degrees of freedom for positioning and grasping. A lightweight end-effector carries the ESP32-CAM, enabling eye-in-hand perception. An external battery with a voltage regulator powers the actuators while maintaining stable controller operation. PWM signals control the servo positions, allowing repeatable movement. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.
The robotic arm employs multiple servo motors providing several degrees of freedom for positioning and grasping. A lightweight end-effector carries the ESP32-CAM, enabling eye-in-hand perception. An external battery with a voltage regulator powers the actuators while maintaining stable controller operation. PWM signals control the servo positions, allowing repeatable movement. The proposed methodology emphasizes robustness,

Corresponding Author: Vaidik Tiwari
Email id: ondutv670@gmail.com
Received: Feb. 2026; Accepted: June 2026; Published online: July 2026

JLISNT
Journal of Leading Innovations in Science and Technology
Journal of Leading Innovations in Science and Technology, Vol.1, No.3, pp:112-131, July 2026.

modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

The robotic arm employs multiple servo motors providing several degrees of freedom for positioning and grasping. A lightweight end-effector carries the ESP32-CAM, enabling eye-in-hand perception. An external battery with a voltage regulator powers the actuators while maintaining stable controller operation. PWM signals control the servo positions, allowing repeatable movement. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

The robotic arm employs multiple servo motors providing several degrees of freedom for positioning and grasping. A lightweight end-effector carries the ESP32-CAM, enabling eye-in-hand perception. An external battery with a voltage regulator powers the actuators while maintaining stable controller operation. PWM signals control the servo positions, allowing repeatable movement. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

The robotic arm employs multiple servo motors providing several degrees of freedom for positioning and grasping. A lightweight end-effector carries the ESP32-CAM, enabling eye-in-hand perception. An external battery with a voltage regulator powers the actuators while maintaining stable controller operation. PWM signals control the servo positions, allowing repeatable movement. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining

Corresponding Author: Vaidik Tiwari
Email id: onduty670@gmail.com

Received: Feb. 2026; Accepted: June 2026; Published online: July 2026

embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

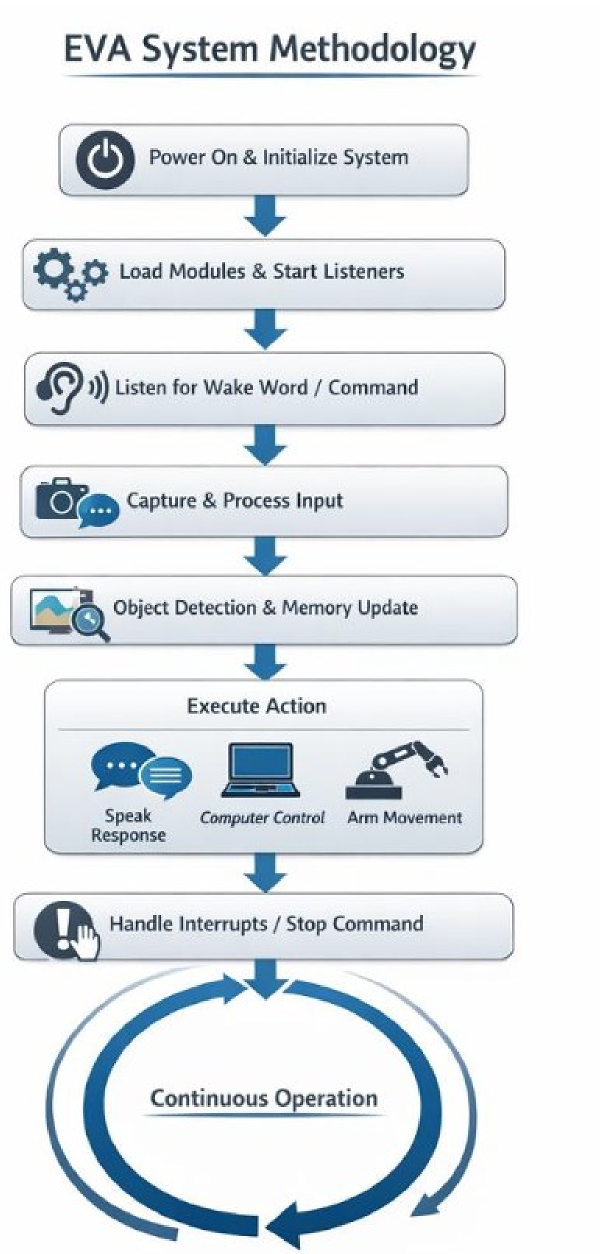

Figure 5. System workflow.

The robotic arm employs multiple servo motors providing several degrees of freedom for positioning and grasping. A lightweight end-effector carries the ESP32-CAM, enabling eye-in-hand perception. An external battery with a voltage regulator powers the actuators

Corresponding Author: Vaidik Tiwari
Email id: ondutv670@gmail.com
Received: Feb. 2026; Accepted: June 2026; Published online: July 2026

while maintaining stable controller operation. PWM signals control the servo positions, allowing repeatable movement. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

IV. Results and Discussion

Experimental evaluation demonstrated reliable speech-based computer control, stable robotic arm movement and successful object detection under indoor lighting. The integrated design reduced command latency and enabled coordinated digital and physical task execution. The ESP32-CAM effectively supported visual observation while spatial memory improved contextual interaction. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

Figure 6: Object Detection using EVA Vision System (Indoor Test)

Corresponding Author: Vaidik Tiwari
Email id: onduty670@gmail.com
Received: Feb. 2026; Accepted: June 2026; Published online: July 2026

JLISNT Journal of Leading Innovations in Science and Technology
Journal of Leading Innovations in Science and Technology, Vol.1, No.3, pp:112-131, July 2026.

Experimental evaluation demonstrated reliable speech-based computer control, stable robotic arm movement and successful object detection under indoor lighting. The integrated design reduced command latency and enabled coordinated digital and physical task execution. The ESP32-CAM effectively supported visual observation while spatial memory improved contextual interaction. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

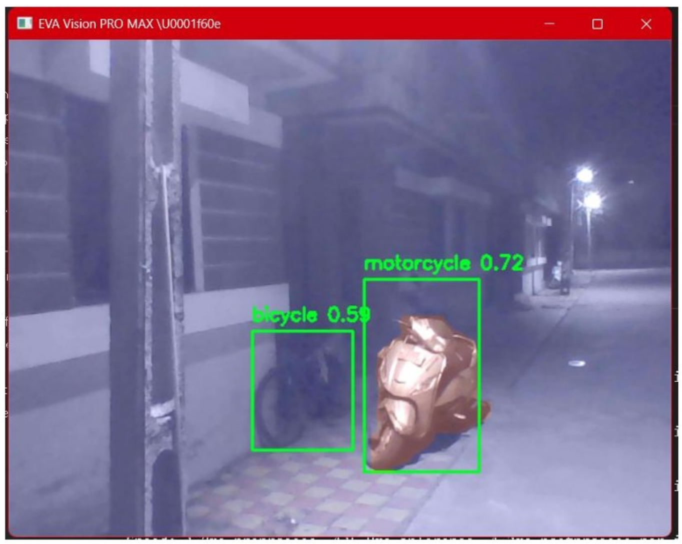

Figure 7: Outdoor Object Detection (Vehicle Identification)

Experimental evaluation demonstrated reliable speech-based computer control, stable robotic arm movement and successful object detection under indoor lighting. The integrated design reduced command latency and enabled coordinated digital and physical task execution. The ESP32-CAM effectively supported visual observation while spatial memory improved contextual interaction. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

Corresponding Author: Vaidik Tiwari
Email id: onduty670@gmail.com

Received: Feb. 2026; Accepted: June 2026; Published online: July 2026

Experimental evaluation demonstrated reliable speech-based computer control, stable robotic arm movement and successful object detection under indoor lighting. The integrated design reduced command latency and enabled coordinated digital and physical task execution. The ESP32-CAM effectively supported visual observation while spatial memory improved contextual interaction. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

Experimental evaluation demonstrated reliable speech-based computer control, stable robotic arm movement and successful object detection under indoor lighting. The integrated design reduced command latency and enabled coordinated digital and physical task execution. The ESP32-CAM effectively supported visual observation while spatial memory improved contextual interaction. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

Experimental evaluation demonstrated reliable speech-based computer control, stable robotic arm movement and successful object detection under indoor lighting. The integrated design reduced command latency and enabled coordinated digital and physical task execution. The ESP32-CAM effectively supported visual observation while spatial memory improved contextual interaction. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

Experimental evaluation demonstrated reliable speech-based computer control, stable robotic arm movement and successful object detection under indoor lighting. The integrated design reduced command latency and enabled coordinated digital and physical

Corresponding Author: Vaidik Tiwari
Email id: onduty670@gmail.com

Received: Feb. 2026; Accepted: June 2026; Published online: July 2026

task execution. The ESP32-CAM effectively supported visual observation while spatial memory improved contextual interaction. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

Experimental evaluation demonstrated reliable speech-based computer control, stable robotic arm movement and successful object detection under indoor lighting. The integrated design reduced command latency and enabled coordinated digital and physical task execution. The ESP32-CAM effectively supported visual observation while spatial memory improved contextual interaction. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

The proposed assistant bridges the gap between virtual assistants and robotic systems by integrating perception, reasoning and action. Although the prototype is intended for research and educational use, the architecture can be extended to industrial inspection, healthcare assistance and smart manufacturing. Future improvements include SLAM, deep learning-based object recognition and autonomous navigation. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

The proposed assistant bridges the gap between virtual assistants and robotic systems by integrating perception, reasoning and action. Although the prototype is intended for research and educational use, the architecture can be extended to industrial inspection, healthcare assistance and smart manufacturing. Future improvements include SLAM, deep learning-based object recognition and autonomous navigation. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency

Corresponding Author: Vaidik Tiwari
Email id: onduty670@gmail.com

Received: Feb. 2026; Accepted: June 2026; Published online: July 2026

response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

The proposed assistant bridges the gap between virtual assistants and robotic systems by integrating perception, reasoning and action. Although the prototype is intended for research and educational use, the architecture can be extended to industrial inspection, healthcare assistance and smart manufacturing. Future improvements include SLAM, deep learning-based object recognition and autonomous navigation. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

The proposed assistant bridges the gap between virtual assistants and robotic systems by integrating perception, reasoning and action. Although the prototype is intended for research and educational use, the architecture can be extended to industrial inspection, healthcare assistance and smart manufacturing. Future improvements include SLAM, deep learning-based object recognition and autonomous navigation. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

The proposed assistant bridges the gap between virtual assistants and robotic systems by integrating perception, reasoning and action. Although the prototype is intended for research and educational use, the architecture can be extended to industrial inspection, healthcare assistance and smart manufacturing. Future improvements include SLAM, deep learning-based object recognition and autonomous navigation. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

Corresponding Author: Vaidik Tiwari
Email id: ondutv670@gmail.com

Received: Feb. 2026; Accepted: June 2026; Published online: July 2026

The proposed assistant bridges the gap between virtual assistants and robotic systems by integrating perception, reasoning and action. Although the prototype is intended for research and educational use, the architecture can be extended to industrial inspection, healthcare assistance and smart manufacturing. Future improvements include SLAM, deep learning-based object recognition and autonomous navigation. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems.

The proposed assistant bridges the gap between virtual assistants and robotic systems by integrating perception, reasoning and action. Although the prototype is intended for research and educational use, the architecture can be extended to industrial inspection, healthcare assistance and smart manufacturing. Future improvements include SLAM, deep learning-based object recognition and autonomous navigation. The proposed methodology emphasizes robustness, modular implementation, efficient communication, low-latency response, and practical deployment in assistive environments. Experimental observations indicate that combining embedded vision, continuous speech interaction, and robotic manipulation improves usability compared with isolated subsystems

V. CONCLUSION

This study presented EVA, an intelligent voice-controlled robotic assistant that integrates speech recognition, computer vision, spatial memory, and robotic manipulation within a unified embedded platform. By combining continuous voice interaction with ESP32-CAM-based visual perception and a multi-degree-of-freedom robotic arm, the system enables seamless coordination between digital automation and physical task execution. The modular hardware and software architecture ensures scalability, low-latency communication, and reliable real-time performance, making the platform suitable for assistive robotics and smart automation applications. Experimental evaluation demonstrated that integrating voice control, object detection, and robotic actuation

Corresponding Author: Vaidik Tiwari

Email id: onduty670@gmail.com

Received: Feb. 2026; Accepted: June 2026; Published online: July 2026

significantly enhances system responsiveness and operational efficiency compared with conventional standalone assistants. Furthermore, the incorporation of spatial memory improves contextual awareness by enabling the system to recognize and recall object locations during interaction. Overall, EVA provides a practical framework for next-generation intelligent service robots and establishes a strong foundation for future research involving autonomous navigation, advanced machine learning, adaptive decision-making, and human-centered robotic assistance..

Author Contribution

Vaidik Tiwari: Conceptualization, methodology, software development, hardware integration, robotic arm design, experimental investigation, data analysis, visualization, and writing original draft preparation.

Jagdish Pampania: Supervision, project administration, conceptual guidance, validation, technical review, manuscript editing, and final approval of the manuscript.

Funding

This research received no external funding

Data Availability Statement

No new data were generated during the study. All the data are contained within the manuscript.

Acknowledgments

We acknowledge the use of Grammarly and Quill Bot, in correcting the English and Grammatical errors in the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

Ethics Declaration

This manuscript is a review article and does not involve any studies with human participants or animals performed by the authors. Therefore, ethical approval and informed consent were not required. The authors declare no conflict of interest, and all referenced works have been properly cited.

References

Corresponding Author: Vaidik Tiwari
Email id: ondutv670@gmail.com
Received: Feb. 2026; Accepted: June 2026; Published online: July 2026

[1]. S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Hoboken, NJ, USA: Pearson, 2021.

[2]. B. Siciliano and O. Khatib, Eds., Springer Handbook of Robotics, 2nd ed. Cham, Switzerland: Springer, 2016.

[3]. R. C. Arkin, Behavior-Based Robotics. Cambridge, MA, USA: MIT Press, 1998.

[4]. L. E. Parker, "Current state of the art in distributed autonomous mobile robotics," in Distributed Autonomous Robotic Systems, Tokyo, Japan: Springer, pp. 3–12, 2000.

[5]. Espressif Systems, ESP32-CAM Technical Reference Manual, Espressif Systems, Shanghai, China, 2023.

[6]. Arduino, "Arduino Uno Rev3 Technical Specifications," Arduino Documentation, 2024. [Online]. Available: https://docs.arduino.cc

[7]. R. Szeliski, Computer Vision: Algorithms and Applications, 2nd ed. Cham, Switzerland: Springer, 2022.

[8]. J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv preprint arXiv:1804.02767, 2018.

[9]. C. Breazeal, Designing Sociable Robots. Cambridge, MA, USA: MIT Press, 2002.

[10]. T. B. Sheridan, "Human–Robot Interaction: Status and Challenges," Human Factors, vol. 58, no. 4, pp. 525–532, 2016..

Corresponding Author: Vaidik Tiwari

Email id: onduty670@gmail.com

Received: Feb. 2026; Accepted: June 2026; Published online: July 2026