International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

JARVIS: A Human Assistant Robot using ROS2 and Computer Vision

Shantanu Hadge, Shweta Mahajan, Asmita Khatavkar, Prof. S. G. Madhikar
Department of Electronics and Telecommunication Engineering Sinhgad College of Engineering, Pune, India

https://doi.org/10.55041/ijsmt.v2i6.093

Cite this Article: Hadge, S., Mahajan, S. & Khatavkar, A. (2026). JARVIS: A Human Assistant Robot using ROS2 and Computer Vision. International Journal of Science, Strategic Management and Technology, *02*(6). https://doi.org/10.55041/ijsmt.v2i6.093

***Abstract*—**Human assistant robots are becoming increasingly important in domestic and workplace environments due to their ability to automate routine tasks and improve human–robot interaction. This paper presents JARVIS, a multifunctional human assistant robot capable of performing voice interaction, hand gesture control, human tracking, object detection, remote operation, obstacle avoidance, and autonomous docking. The system is developed using Raspberry Pi, ESP32, ROS2, OpenCV, MediaPipe, and an offline voice recognition module. Hand ges-tures are recognized using MediaPipe-based landmark detection, while computer vision techniques enable face tracking and object detection. Voice commands allow users to switch between differ-ent operational modes and control robot functions intuitively. The robot employs mecanum wheels for omnidirectional movement and integrates ultrasonic and infrared sensors for navigation and docking. A distributed architecture is adopted in which Raspberry Pi performs high-level processing and ESP32 manages real-time motor control. Experimental results demonstrate suc-cessful implementation of all features with reliable performance under indoor conditions. The proposed system provides a low-cost, scalable, and modular solution for smart assistant and automation applications.

Index Terms—Human Assistant Robot, ROS2, Computer Vi-sion, MediaPipe, Human Tracking, Object Detection, Automatic Docking, Robotics

I. Introduction

The rapid advancement of robotics, artificial intelligence, and embedded systems has accelerated the development of intelligent machines capable of assisting humans in daily activ-ities [1]. Human assistant robots are increasingly being used in domestic, healthcare, educational, and industrial environments to improve convenience, productivity, and safety. These robots combine sensing, perception, decision-making, and actuation capabilities to interact naturally with users and perform a variety of tasks autonomously.

Recent developments in computer vision and machine learning have enabled robots to understand human gestures, recognize objects, track individuals, and respond to voice commands in real time. However, many existing assistant robots are expensive, task-specific, and difficult to deploy in real-world environments. Most systems focus on a single functionality such as object manipulation, navigation, or voice interaction, limiting their practical usability.To address these limitations, this paper presents JARVIS (Just A Rather Very Intelligent System), a multifunctional human assistant robot that integrates multiple intelligent fea-tures within a single platform. The proposed system combines voice control, hand gesture recognition, human tracking, object detection, remote operation, obstacle avoidance, and automatic docking. The robot is built using affordable hardware components including Raspberry Pi, ESP32, Pi Camera, ultrasonic sensors, and a robotic arm mounted on an omnidirectional mobile platform.

ROS2 is utilized as the middleware framework to enable communication between different software modules and hardware components [2]. Computer vision algorithms implemented using OpenCV [3] and MediaPipe [4] provide real-time perception capabilities, while the ESP32 ensures efficient motor control and hardware interfacing. The modular archi-tecture allows easy expansion and integration of additional functionalities.

The main contribution of this work is the development of a low-cost multifunctional assistant robot that combines mobility, manipulation, perception, and human–robot interaction within a single ROS2-based platform. Unlike many existing systems that focus on individual tasks, JARVIS integrates multiple operational modes and autonomous capabilities while maintaining affordability, scalability, and ease of implementation.

The remainder of this paper is organized as follows: Sec-tion II presents the related work. Section III describes the overall system architecture. Section IV discusses the hardware implementation of the proposed robot. Section V explains the methodology adopted for different functionalities. Section VI presents the experimental results and discussion. Finally, Section VII concludes the paper and outlines future research directions.

II. Related Work

Recent advancements in assistive robotics have focused on improving human–robot interaction, autonomous naviga-tion, and intelligent manipulation capabilities [1]. Several researchers have proposed robotic systems capable of assisting users in domestic and professional environments through the integration of computer vision, machine learning, and ad-vanced control techniques.Ayub et al. [5] proposed a personalized household assistive robot capable of learning user preferences and adapting its behavior through continuous human–robot interaction. Their work demonstrated how adaptive learning techniques can improve the effectiveness of robotic assistants in domestic environments through customized task execution.

TABLE I: Comparison of Existing Assistive Robotic Systems

Reference	Main Feature	Limitation
Ayub et al.	Personalized Assis-tance	High Cost
Hagengruber et al.	Daily Activity As-sistance	Complex Sys-tem
VoicePilot	Voice Interaction using LLMs	High Computational Requirement
Huang et al.	Intelligent Manipu-lation using Deep RL	Limited Mobility Features
Proposed JARVIS	Multi-feature Assis-tant Robot	Limited Autonomous Navigation

Hagengruber et al. [1] developed an assistive robotic system designed to support individuals with physical disabilities in performing daily activities. The proposed system utilized a multi-degree-of-freedom robotic manipulator and advanced control strategies to achieve reliable object manipulation and safe human–robot interaction.

Padmanabha et al. introduced VoicePilot, a speech-based robotic interface that utilizes Large Language Models (LLMs) to improve communication between users and assistive robots. The system demonstrated enhanced command understanding and more natural interaction compared to conventional voice-controlled robotic systems.

Huang et al. [6] presented an intelligent robotic arm for assisting elderly individuals using deep reinforcement learning techniques. Their approach improved adaptability and motion planning performance, enabling efficient operation in dynamic environments.

Although these systems demonstrate significant progress in assistive robotics, most are either designed for specific tasks, require expensive hardware, or focus on a single interaction modality. Consequently, there remains a need for a low-cost, multifunctional robotic platform capable of integrating perception, mobility, manipulation, and human–robot interaction within a unified framework.

The proposed JARVIS system addresses this gap by combining voice control, hand gesture recognition, human tracking, object detection, obstacle avoidance, remote operation, and autonomous docking within a single ROS2-based architecture using affordable hardware components.

III. System Architecture

The proposed JARVIS system is designed as a multifunctional human assistant robot capable of performing voice interaction, gesture-based control, human tracking, object detection, remote operation, obstacle avoidance, and autonomous docking. The system integrates perception, decision-making, and motion control modules within a unified ROS2-based framework.The overall architecture consists of three major layers: sensing, processing, and actuation. The sensing layer includes the Pi Camera, ultrasonic sensors, infrared sensors, and the VC02 voice recognition module. These devices continuously collect environmental and user interaction data.

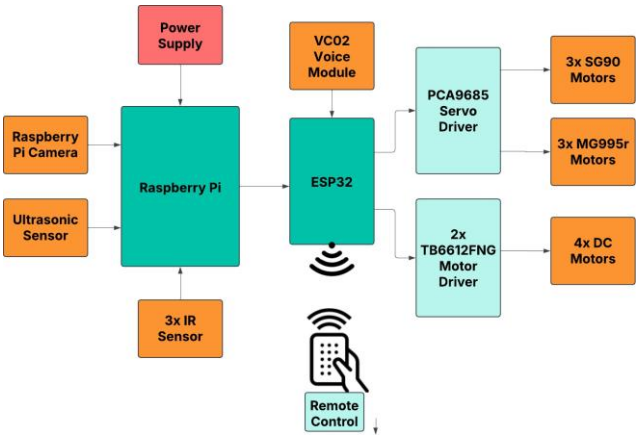


Fig. 1: Overall System Architecture of JARVIS

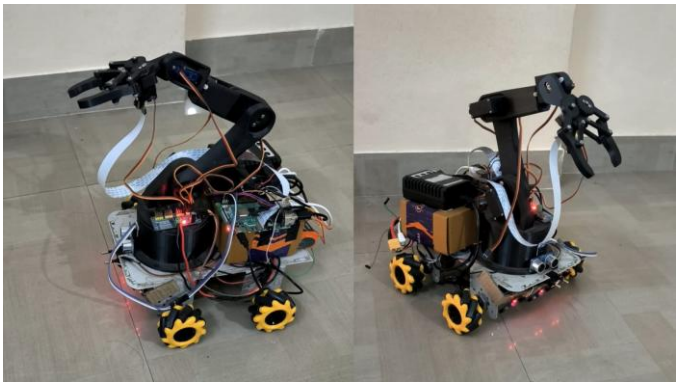


Fig. 2: Developed JARVIS Human Assistant Robot

The processing layer is composed of a Raspberry Pi and an ESP32 microcontroller. The Raspberry Pi performs computationally intensive tasks such as computer vision, object detection, face tracking, gesture recognition, and ROS2 node management. The ESP32 is responsible for real-time motor control and sensor interfacing.

The actuation layer consists of the robotic arm, DC motors, servo motors, and mecanum wheel drive system. Based on commands generated by the processing layer, the actuators execute movements required for navigation, object manipulation, and user interaction.

ROS2 serves as the communication backbone of the sys-tem and enables efficient communication between software modules [2].Each functionality, including voice control, ges-ture recognition, human tracking, object detection, remote operation, and docking, is implemented as an independent ROS2 node. This modular architecture improves scalability, maintainability, and future extensibility of the system.

The combination of affordable hardware components and modular software architecture enables the robot to perform multiple intelligent tasks while maintaining low implementation cost and reliable real-time performance.

International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

IV. Hardware Implementation

The JARVIS robot is developed using a combination of embedded computing platforms, sensing devices, actuation systems, and communication modules. The hardware architecture is designed to provide reliable real-time performance while maintaining low implementation cost and modularity.

Fig. 3: Mechanical Structure and Components of the Robotic Arm

The robotic arm consists of six rotational joints that provide multiple degrees of freedom for object manipulation. Individual servo motors actuate each joint, enabling shoulder, elbow, wrist, and gripper movements. The joint arrangement allows the arm to perform pick-and-place operations and gesture-based motion control while maintaining a compact mechanical structure.

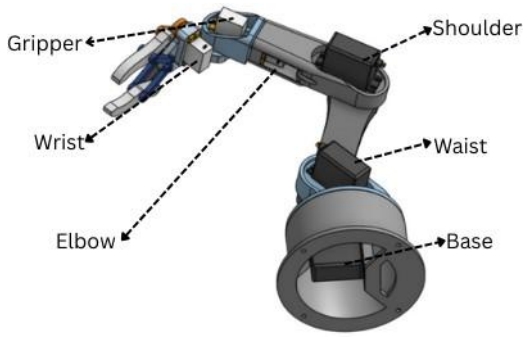


Fig. 4: Joint Configuration of the 6-DOF Robotic Arm

The primary processing unit of the system is a Raspberry Pi, which performs computationally intensive tasks including computer vision, gesture recognition, human tracking, object detection, and ROS2 node management. An ESP32 microcontroller is employed for low-level control operations such as motor driving, sensor interfacing, and actuator control. Communication between the Raspberry Pi and ESP32 enables efficient distribution of computational and control tasks.

A Pi Camera is utilized as the primary perception sensor for image acquisition. The camera provides visual input for gesture recognition, face tracking, object detection, and ArUco marker-based docking. Voice interaction is implemented using the VC02 offline voice recognition module, allowing the robot to respond to predefined commands without requiring internet connectivity.

Mobility is achieved through a mecanum wheel-based drive system that enables omnidirectional movement. The platform can move forward, backward, sideways, and diagonally, thereby improving maneuverability in indoor environments. Ultrasonic sensors and infrared sensors are integrated for obstacle avoidance and docking assistance.

The robot is also equipped with a six-degree-of-freedom robotic arm capable of performing object manipulation tasks. Multiple servo motors provide rotational movement for indi-vidual joints, enabling pick-and-place operations and gesture-based control. The arm is mounted on the mobile platform, allowing the robot to combine mobility and manipulation capabilities within a single system.

The combination of embedded processing, computer vision, omnidirectional mobility, and robotic manipulation enables the proposed system to perform a wide range of assistive and autonomous tasks. The modular hardware design further allows future integration of additional sensors and intelligent functionalities.

V. Methodology

The proposed JARVIS system operates through a com-bination of computer vision, voice interaction, autonomous navigation, and robotic manipulation techniques. Different operational modes are implemented as independent ROS2 nodes, enabling modular execution and efficient communica-tion between system components.

A. Voice Control

Voice interaction is implemented using the VC02 offline voice recognition module. Predefined voice commands are used to activate different operational modes such as gesture control, human tracking, object detection, remote operation, and automatic docking. The recognized command is transmit-ted to the ESP32 and corresponding ROS2 nodes are activated. This enables intuitive and hands-free interaction between the user and the robot.

B. Human Tracking

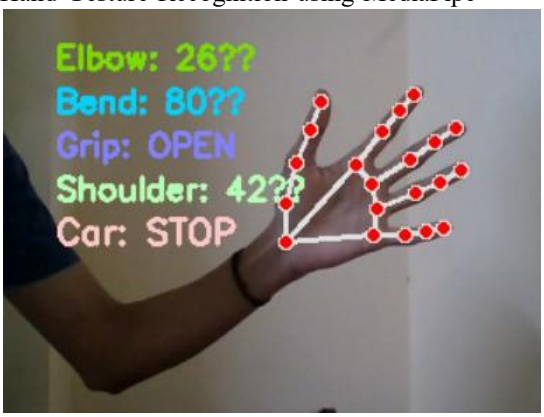
The human tracking module enables the robot to detect and follow a user in real time. The Pi Camera continuously captures video frames, which are processed using computer vision algorithms to detect a human face. The position of the detected face is compared with the center of the image

frame and motion commands are generated to align the robot with the user. Once aligned, the robot follows the user while maintaining a safe distance.

C. Hand Gesture Control

Hand gesture recognition is implemented using MediaPipe [4]. The framework detects 21 hand landmarks from the camera feed and calculates their relative positions. Different gestures are mapped to robotic arm and mobile base movements. Finger count, hand position, and hand orientation are used to control navigation and manipulation tasks in real time.

Fig. 5: Hand Gesture Recognition using MediaPipe



D. Object Detection

Object detection is performed using computer vision techniques executed on the Raspberry Pi using concepts derived from modern object detection frameworks [7], [8]. The camera continuously scans the environment and identifies target ob-jects based on trained detection models. After detection, the robot aligns itself with the object and moves toward it. This functionality enables environmental awareness and supports future object manipulation tasks.

E. Automatic Docking

The docking system combines ArUco marker detection [9] and line-following techniques. When docking is initiated, the robot searches for a predefined ArUco marker placed near the charging station. After detecting and aligning with the marker, the robot approaches the docking area. Infrared sensors are then used to follow a guiding line for accurate final positioning at the charging station.

The integration of these modules allows the robot to per-form multiple intelligent tasks while maintaining reliable real-time operation. The ROS2-based architecture ensures seamless communication between all functionalities and enables easy expansion of future features.

VI. Results and Discussion

The proposed JARVIS system was successfully implemented and tested under indoor operating conditions. Individual modules were initially evaluated independently and later

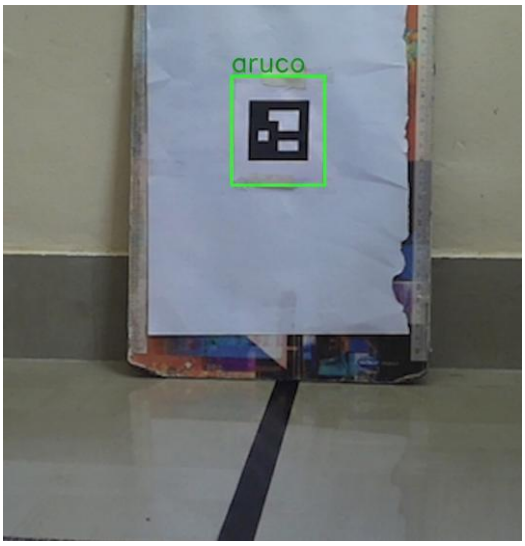


Fig. 6: ArUco Marker-Based Automatic Docking

integrated into a unified ROS2-based framework. Experimental evaluation demonstrated the robot's capability to perform voice interaction, gesture-based control, human tracking, ob-ject detection, obstacle avoidance, and automatic docking in real time.

The voice recognition module successfully identified predefined commands and enabled seamless switching between different operational modes. The offline VC02 module provided reliable performance without requiring internet connectivity. Similarly, the gesture recognition module accurately detected hand landmarks using MediaPipe and enabled intuitive control of the robotic arm and mobile platform.

The human tracking module demonstrated stable perfor-mance under normal indoor lighting conditions. The robot was able to continuously detect, track, and follow a user while maintaining alignment and a safe operating distance.

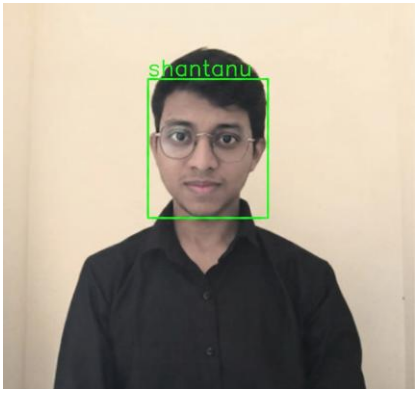


Fig. 7: Human Tracking and Face Detection Results

Object detection experiments showed successful identifica-tion and localization of target objects using computer vision techniques. The robot was able to recognize objects in real

time and align itself with the detected target, demonstrating effective environmental perception capabilities.

Fig. 8: Real-Time Object Detection Results

The automatic docking module effectively combined ArUco marker detection with infrared sensor-based line following. The robot successfully detected the docking station, aligned itself with the marker, and completed the docking process with acceptable positioning accuracy.

ROS2 enabled efficient communication between software modules and simplified overall system integration. The modular architecture allowed independent execution of different functionalities while maintaining reliable data exchange between nodes.

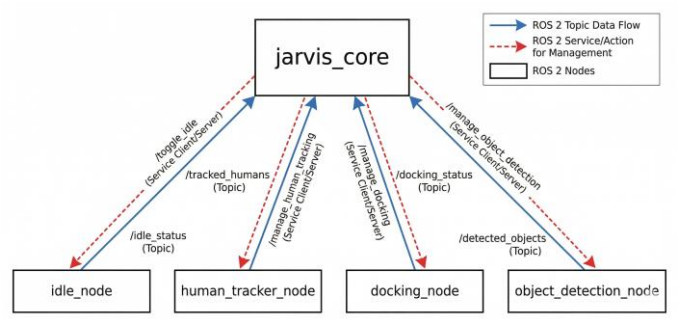


Fig. 9: ROS2 Node Communication Architecture

The experimental results indicate that the proposed system achieved reliable performance across all implemented functionalities. Voice command recognition achieved an accuracy of approximately 85%, while hand gesture recognition using MediaPipe achieved an accuracy of 90% under normal indoor lighting conditions. Human tracking was successfully performed within a range of 0.5–2.5 meters. Object detec-tion achieved approximately 80% accuracy with an average processing speed of 12 frames per second on the Raspberry Pi platform. The autonomous docking system successfully completed 8 out of 10 docking trials, demonstrating reliable positioning performance.

Although the developed system demonstrated satisfactory performance, certain limitations were observed. Vision-based functionalities were influenced by lighting conditions, while computationally intensive processing tasks introduced minor

TABLE II: Experimental Performance Evaluation

Parameter	**Observed Performance**
Voice Command Accuracy	85%
Hand Gesture Recognition Accuracy	90%
Human Tracking Range	0.5–2.5 m
Object Detection Accuracy	80%
Average Object De-tection FPS	12 FPS
Docking Success Rate	8/10 Trials
Obstacle Detection Range	10–150 cm
ROS2 Communica-tion Delay	< 500 ms

delays. Despite these limitations, the proposed system successfully achieved the intended objectives and validated the feasibility of integrating multiple intelligent functionalities within a single low-cost robotic platform.

VII. Conclusion and Future Work

This paper presented JARVIS, a multifunctional human assistant robot designed to assist users through intelligent interaction, autonomous operation, and robotic manipulation. The developed system successfully integrates voice control, hand gesture recognition, human tracking, object detection, remote operation, obstacle avoidance, and automatic docking within a unified ROS2-based architecture.

The proposed system combines Raspberry Pi, ESP32, computer vision techniques, and embedded control systems to achieve reliable real-time performance. MediaPipe-based gesture recognition enabled intuitive robotic arm control, while face tracking and object detection enhanced the robot's perception capabilities. The implementation of ROS2 provided efficient communication between software modules and simplified system integration. Experimental results demonstrated the successful operation of all implemented functionalities, validating the effectiveness of the proposed architecture for smart assistant applications.

The developed prototype demonstrates that multiple intelli-gent features can be integrated into a single low-cost robotic platform while maintaining modularity, scalability, and ease of implementation. The system can be utilized in domestic environments, educational robotics, research applications, and smart automation systems.Future work will focus on improving the intelligence, au-tonomy, and manipulation capabilities of the robot. Simul-taneous Localization and Mapping (SLAM) techniques [10] using LiDAR and depth sensing will be integrated to enable autonomous room mapping, localization, and navigation in dynamic environments. The robot will be enhanced with an AI-powered voice assistant based on Large Language Models (LLMs) to support natural language conversations, contextual

understanding, task planning, and personalized user interac-tion.

Further improvements include the development of a stronger robotic arm with higher payload capacity for handling heavier objects and performing complex manipulation tasks. Advanced path planning and obstacle avoidance algorithms will be implemented to improve navigation efficiency. Integration of machine learning and reinforcement learning techniques will enable adaptive decision-making and autonomous task exe-cution. Additional enhancements such as cloud connectivity, IoT integration, multi-room navigation, object grasping opti-mization, and intelligent home automation support will further extend the capabilities of the proposed system and move it closer to a fully autonomous personal assistant robot.

REFERENCES

[1] B. Siciliano and O. Khatib, *Springer Handbook of Robotics*, 2nd ed. Springer, 2016.
[2] S. Macenski, F. Martin, R. White, and J. Clavero, "The robot operating system 2 (ros 2): Design, architecture, and uses in the wild," *Science Robotics*, vol. 7, no. 66, 2022.
[3] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Media, 2021.
[4] C. Lugaresi, J. Tang, H. Nash *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.
[5] S. Thrun, W. Burgard, and D. Fox, "Probabilistic robotics," *MIT Press*, 2005.
[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Informa-tion Processing Systems*, vol. 25, 2012.
[7] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Op-timal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
[9] S. Garrido-Jurado, R. Munoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marin-Jimenez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6,
pp. 2280–2292, 2014.
[10] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira,
I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.
[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of CVPR*, 2001, pp. 511–518.
[12] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schul-man, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.