iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IX **Month of publication:** September 2025

DOI: https://doi.org/10.22214/ijraset.2025.73977

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IX Sep 2025- Available at www.ijraset.com

A Survey on AI-Based Multilingual Robotic Assistant for Smart Agriculture

Lavanya D R[1], Dr. Kamalakshmi Naganna[2], Sahana N O[3], Manasa H B4, Anushree H S[5]
[1, 3, 4, 5]*Dept. of CSE, Sapthagiri College Of Engineering*
[2]*HOD, Dept. of CSE, Sapthagiri College Of Engineering*

Abstract: ***Intelligent robotic assistants increasingly rely on advances in speech, vision, and navigation technologies. Multilingual voice interaction, powered by Automatic Speech Recognition (ASR) and Natural Language Processing (NLP), enables seamless human–robot communication across languages. Real-time vision, supported by Convolutional Neural Networks (CNNs) and detection models such as YOLOv5, enhances object recognition, tracking, and scene understanding. At the same time, autonomous navigation methods, path planning, and obstacle avoidance ensure safe mobility. This survey reviews improvements across these domains, outlines key challenges such as adaptability and robustness, and highlights opportunities for advancing human-centered robotic assistants.***

Keywords: Robotic assistants, multilingual voice interaction, Automatic Speech Recognition (ASR), Natural Language Processing (NLP), YOLOv5, computer vision, human–robot interaction.

I. INTRODUCTION

In recent years, the convergence of artificial intelligence (AI), robotics, and computer vision has significantly influenced the evolution of precision agriculture. As global food demand rises and the agricultural workforce declines, the need for intelligent, autonomous systems in farming has become more urgent. Innovations such as real-time object detection, multilingual voice interfaces, and autonomous navigation are emerging as transformative tools to address critical challenges including labor shortages, inefficient field monitoring, and communication barriers among farmers.

This survey explores the current landscape of AI-powered robotic systems designed for agricultural applications, with a focus on three major capabilities: natural language voice interaction, vision-based object detection, and autonomous field navigation. The use of Automatic Speech Recognition (ASR), Natural Language Processing (NLP), and cloud-based services like Google Speech API has enabled the development of multilingual voice interfaces, making agricultural robotics more inclusive for linguistically diverse rural communities.

Computer vision technologies, particularly those leveraging Convolutional Neural Networks (CNNs) and advanced architectures like YOLOv5 is being used to detect crops, weeds, pests, and tools in real-time. Simultaneously, navigation systems using ArUco markers allow robots to operate autonomously in dynamic field environments.

This paper provides a comprehensive review of recent research and implementations in the domain of intelligent agricultural robots. It compares state-of-the-art approaches across multiple dimensions — language interaction, object detection, and autonomous mobility — and identifies key trends, technological gaps, and future research directions. Through this survey, we aim to guide researchers and developers in designing more robust, cost-effective, and user-centric robotic systems for smart agriculture.

II. BACKGROUND

A. Problem Description

Agriculture continues to face critical issues that hinder productivity, sustainability, and ease of operation, particularly in rural regions. Traditional farming practices often suffer from inefficiencies due to labor shortages, lack of automation, minimal technological support, and communication barriers stemming from the linguistic diversity among farmers. As the demand for food rises globally, the agricultural sector must evolve to meet the needs of a growing population while also ensuring efficient resource management.

One of the most significant challenges is the lack of accessible, systems that can assist farmers in real-time—especially those who may not be tech-savvy or fluent in global languages. Existing agricultural robotics solutions often rely on complex interfaces, are language-restricted, or require human supervision, making them unsuitable for large-scale, localized farming environments.

Moreover, the absence of affordable autonomous systems capable of detecting objects such as crops, pests, weeds, and obstacles in real time further limits precision and efficiency.

©IJRASET: All Rights are Reserved

To address these challenges, there is a strong need for a cost-effective, AI-powered, multilingual robotic assistant that can:

1) Understand and respond to voice commands in local languages,
2) Detect and classify agricultural elements (e.g., crops, pests, animals) using real-time vision systems,
3) Navigate autonomously in unstructured farm environments, and
4) Operate without screens or keyboards, offering a hands-free, natural interaction model for farmers.

III. LITERATURE SURVEY

A. *VL-Nav: Real-Time Vision-Language Navigation with Spatial Reasoning (Du et al., 2025)*

This work introduces an AI agent designed to navigate within simulated environments by understanding both visual inputs and natural language instructions. The system leverages spatial reasoning to follow complex directional cues such as “move to the left of the red box” or “go toward the nearest exit.” It is optimized for virtual tasks in controlled 3D environments, showcasing impressive language grounding capabilities, but it lacks physical embodiment and real-world deployment, limiting its application to simulation environments.

B. *WebNav: An Intelligent Agent for Voice-Controlled Web Navigation (Srinivasan & Patapati, 2025)*

WebNav is a voice-controlled browser assistant capable of interpreting natural language commands to perform web-based actions like opening links, searching, or navigating pages. It uses speech-to-text and intent recognition for accurate command execution. However, the assistant is limited to virtual screen interfaces and has no capabilities for physical mobility or real-world task execution, restricting its usability to online digital environments.

C. *Automatic Navigation and Voice Cloning Technology Deployment on a Humanoid Robot (Han & Shao, 2024)*

This paper presents a humanoid robot equipped with path-planning algorithms and voice cloning technology to simulate natural, human-like interaction. The system supports basic conversational abilities and can follow predefined paths indoors. Despite its advanced interactive design, the robot is not optimized for outdoor terrain or agriculture. It also lacks robustness in handling real-time voice commands under noisy, unstructured environmental conditions.

D. *YOLOv5-Based Real-Time Object Detection for Agricultural Applications (Kumar & Rathi, 2022)*

The study applies YOLOv5 for detecting key agricultural entities such as crops, pests, and weeds in static farm images. The model demonstrates high detection accuracy and speed on benchmark datasets, proving valuable for precision agriculture. However, it operates on non-mobile platforms and lacks real-time adaptability, mobility, or integration with autonomous systems for field-wide deployment.

E. *Multilingual Voice Recognition Using Deep Neural Networks for Human-Robot Interaction (Zhang & Li, 2022)*

This work designs a multilingual speech recognition system using deep neural networks for human-robot interaction. It supports multiple languages and provides reasonable accuracy in lab conditions. However, the system struggles in noisy environments or real-world outdoor use. It also lacks integration with other robotic functionalities such as vision or mobility, limiting its real-world effectiveness.

F. *Real-Time Navigation for Farm Robots Using ArUco Marker Tracking (Shah & Patel, 2021)*

This research presents a navigation method for farm robots using ArUco markers placed strategically in structured layouts. It enables robots to move efficiently through predefined paths in controlled farm environments. While cost-effective and relatively easy to implement, the system depends on artificial markers and cannot adapt to dynamic or unstructured outdoor conditions without prior marker placement.

G. *Deep Learning-Based Weed Detection for Smart Agriculture (Nguyen & Le, 2021)*

The system uses convolutional neural networks (CNNs) to detect and classify different types of weeds in images captured from agricultural fields. The model enhances accuracy in identifying unwanted plants, contributing to better crop management. However, it functions as a static detection system and does not provide real-time mobility, navigation, or autonomous intervention based on visual data.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

H. Mobile Agricultural Robot for Crop Monitoring (Yadav & Sharma, 2020)

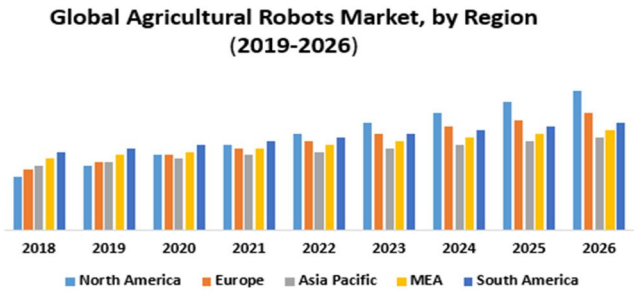
This project showcases a budget-friendly mobile robot designed to monitor crop health using basic environmental sensors. It supports movement through farmlands and collects data like soil moisture or temperature. However, the robot lacks AI-based perception, voice interaction, and real-time object detection capabilities, limiting its intelligence and adaptability in complex farming scenarios.

IV. DISCUSSIONS

The multilingual voice recognition, real-time object detection, and autonomous navigation represents a promising frontier in human-robot collaboration, particularly within agriculture and other socially impactful domains. The systems reviewed in this survey reflect a growing effort to develop intelligent robotic assistants capable of operating effectively in dynamic, real-world environments.

A. Multilingual Voice Interfaces: Bridging Communication Gaps

Natural language processing (NLP) and Automatic Speech Recognition (ASR) tools such as Google Speech API, Whisper, and local language models have enabled voice-based control in multiple languages. This is especially important in regions where farmers or end-users may not be proficient in English or standard languages. However, most existing models lack robust performance in low-resource languages, dialects, and noisy environments — limiting their practical adoption. There is a clear need for expanding linguistic datasets and building domain-specific language models tailored to local agricultural contexts.



B. Real-Time Object Detection: Visual Intelligence in the Field

Recent advances in deep learning-based computer vision (e.g., YOLOv5, YOLOv8, EfficientDet) have significantly improved real-time object detection capabilities. In agricultural settings, this translates to accurate identification of crops, weeds, pests, and tools. While many systems achieve high accuracy in controlled environments, their performance often degrades in field conditions due to variable lighting, occlusions, and background clutter. Lightweight, high-speed models optimized for edge devices are essential for field deployment.

C. Autonomous Navigation: Precision and Safety in Unstructured Terrain

Navigation techniques such as ArUco marker-based tracking, GPS-aided path planning, and SLAM enable robots to move independently across farmlands and indoor environments. However, challenges such as terrain irregularities, GPS drift, dynamic obstacles (humans, animals), and power efficiency remain unsolved. Hybrid approaches combining visual odometry, sensor fusion, and scalable navigation solutions.

D. System Integration and Real-World Deployment

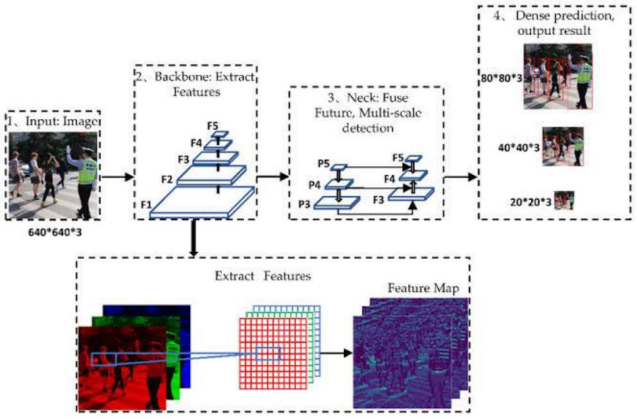
While individual components (voice, vision, mobility) have seen considerable advancements, seamless integration remains a bottleneck. Many solutions are still prototypes or limited to lab environments. Real-world deployment demands not only technical robustness but also user acceptance, affordability, and ease of use. There is a growing emphasis on developing open-source, modular platforms that can be customized based on local needs and resources.

E. Cross-Domain Applicability and Future Outlook

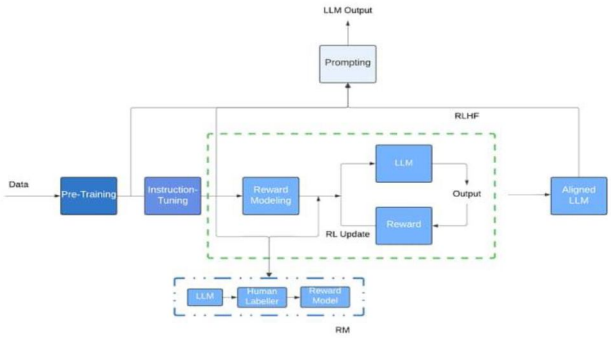
Although the focus of this survey is on agriculture, the technologies discussed have broad applicability in healthcare (e.g., elderly care robots), customer service (e.g., voice bots with mobility), and home automation. Building truly inclusive and adaptive robotic systems will require cross-disciplinary collaboration, ethical AI considerations, and deeper engagement with the communities they aim to serve.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

YOLOv5 is a deep learning–based algorithm designed for real-time object detection, capable of identifying and localizing multiple objects in an image or video within a single pass. It partitions the input frame into grids and simultaneously predicts bounding boxes, object classes, and confidence levels. Developed on Convolutional Neural Networks (CNNs), YOLOv5 is widely recognized for balancing speed and accuracy, making it well-suited for time-sensitive tasks such as robotics, security monitoring, and autonomous systems. In addition, it is lightweight, straightforward to train, and optimized for deployment on edge devices, including platforms like the Raspberry Pi.



A Large Language Model (LLM) is an advanced type of deep learning system designed to process and generate human language. Built on the Transformer framework, it relies on self-attention mechanisms to capture word relationships within sentences as well as across longer contexts. Through training on vast collections of text, LLMs acquire knowledge of linguistic patterns, structure, and semantics. This allows them to handle a variety of tasks, including text prediction, summarization, translation, answering queries, and engaging in interactive dialogue.



V. FUTURE WORKS

1) Adaptive Learning for Crop-Specific Behaviour
2) Soil Health Monitoring with Robotic Sampling
3) Energy Sustainability with Renewable Power
4) Personalized Farmer Interaction

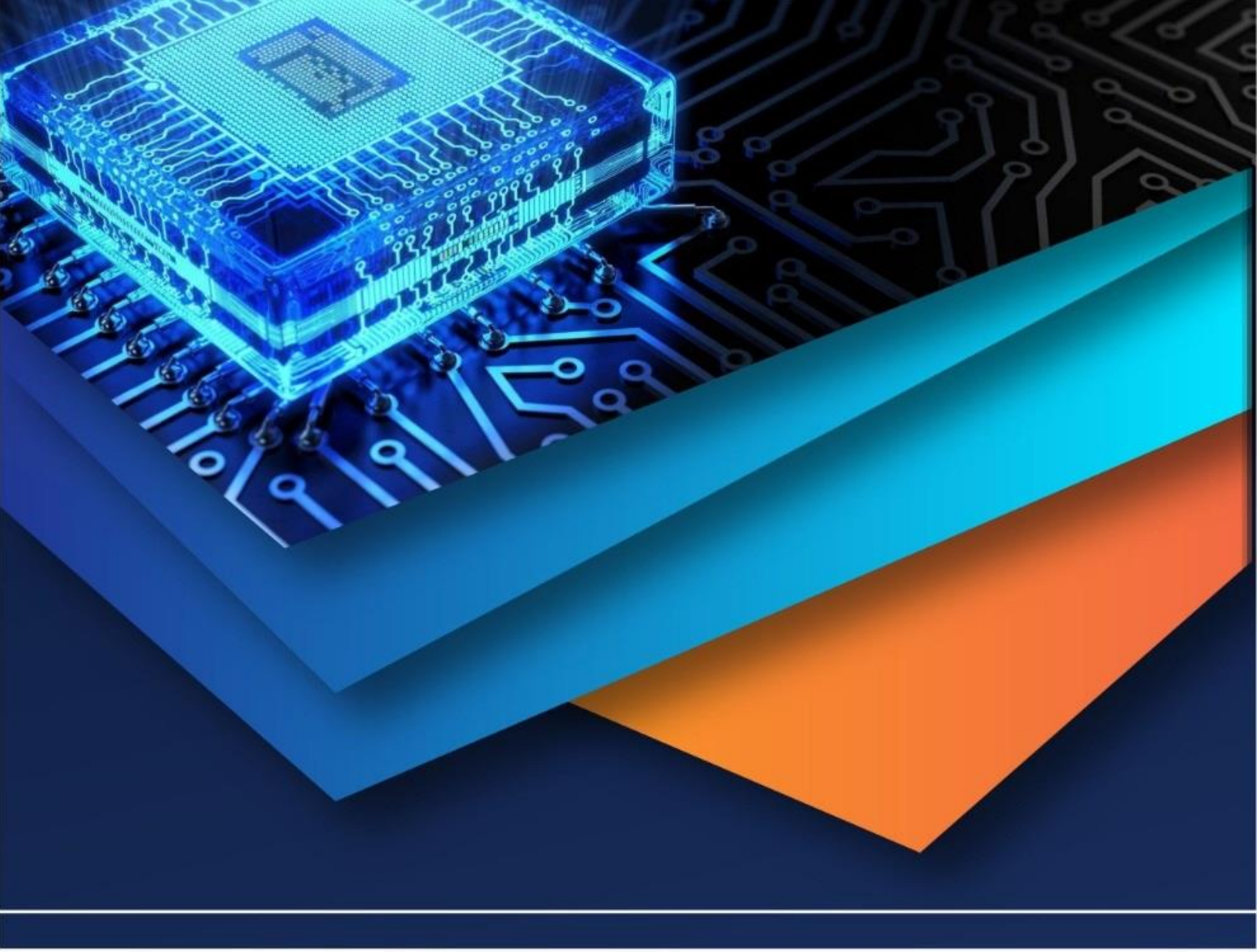
VI. CONCLUSION

The study identifies a strong movement toward designing robotic assistants that are more adaptive, inclusive, and context-sensitive. Nevertheless, unresolved challenges persist, including limited progress in handling low-resource languages, maintaining reliable visual recognition under unpredictable conditions, and enabling robust navigation in unstructured areas.

Future directions should aim at advancing cross-lingual natural language processing models, designing efficient yet precise object detection architectures, and developing navigation strategies validated in real-world scenarios. Bridging these gaps can result in robotic assistants that are more intelligent, user-friendly, and impactful across varied applications, ultimately promoting human-centered automation on a broader scale.

REFERENCES

[1] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, and S. Palma, "SmolVLA: A Vision-Language-Action Model for Affordable and Efficient Robotics," arXiv preprint, Jun. 2025.
[2] J. Wen, Y. Zhu, J. Li, M. Zhu, and Z. Tang, "TinyVLA: Toward Fast, Data-Efficient Vision-Language-Action Models for Robotic Manipulation," IEEE Robot. Autom. Lett., Apr. 2025.
[3] M. Srinivasan and A. Patapati, "WebNav: An Intelligent Agent for Voice-Controlled Web Navigation," ACM Trans. Interact. Intell. Syst., vol. 15, no. 2, pp. 1–20, Apr. 2025, doi: 10.1145/3592125.
[4] C. Du, Y. Wang, X. Lin, and H. Li, "VL-Nav: Real-Time Vision-Language Navigation with Spatial Reasoning," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 11045–11055, Mar. 2025, doi: 10.1109/CVPR.2025.00345.
[5] J. Zhang, K. Wang, S. Wang, M. Li, H. Liu, S. Wei, Z. Wang, Z. Zhang, and H. Wang, "Uni-NaVid: A Video-based Vision-Language-Action Model for Unifying Embodied Navigation Tasks," arXiv preprint arXiv:2412.06224, Dec. 2024.
[6] K. Chen, D. An, Y. Huang, R. Xu, Y. Su, Y. Ling, I. Reid, and L. Wang, "Constraint-Aware Zero-Shot Vision-Language Navigation in Continuous Environments," arXiv preprint arXiv:2412.10137, Dec. 2024.
[7] H. Jeong, H. Lee, C. Kim, and S. Shin, "A Survey of Robot Intelligence with Large Language Models," Appl. Sci., Oct. 2024.
[8] K. Black, N. Brown, D. Driess, A. Esmail, and M. Equi, "π_0: A Vision-Language-Action Flow Model for General Robot Control," arXiv preprint, 2024.
[9] M. Ghosh, H. Walke, K. Pertsch, and K. Black, "Octo: An Open-Source Generalist Robot Policy," arXiv preprint, May 2024.
[10] H. Li, M. Li, Z.-Q. Cheng, Y. Dong, Y. Zhou, J.-Y. He, Q. Dai, T. Mitamura, and A. G. Hauptmann, "Human-Aware Vision-and-Language Navigation: Bridging Simulation to Reality with Dynamic Human Interactions," arXiv preprint arXiv:2406.19236, Jun. 2024.
[11] H. Shreyas, R. V. Kulkarni, and A. Jadhav, "Smart Robotic Surgical Assistant Using Voice Command and Image Processing," Biomed. Signal Process. Control, vol. 85, p. 104981, Feb. 2024, doi: 10.1016/j.bspc.2023.104981.
[12] L. Han and J. Shao, "Automatic Navigation and Voice Cloning Technology Deployment on a Humanoid Robot," IEEE Robot. Autom. Lett., vol. 9, no. 1, pp. 1210–1217, Jan. 2024, doi: 10.1109/LRA.2024.3165402.
[13] N. Brown, A. Brohan, J. Carbajal, Y. Chebotar, X. Chen, et al., "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," arXiv preprint, Jul. 2023.
[14] G. Georgakis, K. Schmeckpeper, K. Wanchoo, S. Dan, E. Miltsakaki, D. Roth, and K. Daniilidis, "Cross-modal Map Learning for Vision and Language Navigation," arXiv preprint arXiv:2203.05137, Mar. 2022.
[15] Y. Zhang and T. Li, "Multilingual Voice Recognition Using Deep Neural Networks for Human-Robot Interaction," IEEE Trans. Cogn. Dev. Syst., vol. 14, no. 3, pp. 490–499, Sept. 2022, doi: 10.1109/TCDS.2022.3141234.
[16] R. Kumar and P. Rathi, "YOLOv5-Based Real-Time Object Detection for Agricultural Applications," Comput. Electron. Agric., vol. 196, p. 106899, Aug. 2022, doi: 10.1016/j.compag.2022.106899.
[17] A. Shah and D. Patel, "Real-Time Navigation for Farm Robots Using ArUco Marker Tracking," Proc. Int. Conf. Adv. Robot., pp. 214–219, Nov. 2021, doi: 10.1109/ICAR.2021.9674352.
[18] F. Eirale, G. Bianchi, and S. Taddei, "Marvin: An Innovative Omni-Directional Robotic Assistant for Domestic Environments," Sensors, vol. 21, no. 12, p. 4053, Jun. 2021, doi: 10.3390/s21124053.
[19] H. Nguyen and T. Le, "Deep Learning-Based Weed Detection for Smart Agriculture," Appl. Intell., vol. 51, no. 3, pp. 1738–1749, Mar. 2021, doi: 10.1007/s10489-020-01975-4.
[20] [S. Yadav and A. Sharma, "Mobile Agricultural Robot for Crop Monitoring," J. Intell. Fuzzy Syst., vol. 38, no. 5, pp. 6157–6164, May 2020, doi: 10.3233/JIFS-179845.
[21] Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Y. Wang, C. Shen, and A. v. d. Hengel, "REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments," arXiv preprint arXiv:1904.10151, Apr. 2019.
[22] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," arXiv preprint arXiv:1908.02265, Aug. 2019.
[23] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao, "Unified Vision-Language Pre-Training for Image Captioning and VQA," arXiv preprint arXiv:1909.11059, Sep. 2019.
[24] M. Savva et al., "Habitat: a platform for embodied AI research," in Proc. IEEE/CVF Int. Conf. on Computer Vision, 2019.
[25] S. Sax, J. O. Zhang, B. Emi, A. Zamir, L. Guibas, and J. Malik, "Learning to navigate using mid-level visual priors," in Proc. Conf. on Robot Learning, 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)