Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD28956

Application of Autonomous Decision-Making Multimodal Perception Systems in Different Fields

Weijia Hu

Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, China
12411824@sustech.edu.cn

Abstract. In the current context, as robotic technology penetrates deeper into more complex scenarios, autonomous decision-making capability has become a core indicator for evaluating robot performance. The multimodal perception system, as the central hub for robots to acquire environmental information and understand task scenarios, this paper explores its applications in different fields and analyzes the existing challenges. Combining specific cases from the past five years, it systematically analyzes the application modes of multimodal perception systems in three typical robotic autonomous decision-making fields: agriculture, bionics, and medical care. It concludes that the current multimodal perception systems face core challenges at both technical and safety-ethical levels. Finally, it summarizes the research limitations of this paper in terms of extreme environments and lightweight design, and proposes future development directions. The findings reveal that advancing multimodal perception systems is essential for enhancing autonomous decision-making in robots, yet addressing the identified technical and safety-ethical challenges will be crucial for successful implementation across diverse applications.

Keywords: Multimodal perception system, Robot, Principle, Application.

1. Introduction

Imitating human sensory perception, gaining in-depth understanding of complex task environments, and quickly executing high-difficulty tasks are among the unremitting pursuits in the field of robotics. Sensors play a crucial role in this endeavor, enabling robots to perceive, evaluate, and interact with their surroundings. As intelligent robots evolve toward greater flexibility and miniaturization, there is a growing need for integrated and adaptable components. This requires sensors to be compatible with various signal processing technologies and easily integrated into small devices. Since single-function sensors are not conducive to achieving the development goals of multi-functional integration, flexibilization, and miniaturization, multimodal sensors—integrated sensors with multiple functions such as electricity, optics, and magnetism—can endow robots with stronger perception capabilities, more complex task execution capabilities, and more intelligent learning mechanisms. They also feature high resolution, high sensitivity, and fast response, leading to a surge in interest in their applications.

DOI: 10.54254/2755-2721/2025.LD28956

This paper reviews the definition and principles of multimodal perception systems, focuses on their applications in different robotic fields, and predicts the future wide application of multimodal perception systems. Additionally, it addresses existing technical challenges and safety-ethical issues, providing a reference for subsequent improvements of the system.

2. Theoretical overview

A multimodal perception system integrates various sensing modules, including vision, touch, sound, and environmental parameters, to collect, preprocess, and fuse multi-dimensional information. This system enhances environmental cognition and decision-making for intelligent agents like robots. Its primary advantage is reducing uncertainty in single-modal perception, thereby improving accuracy in understanding complex scenarios and supporting autonomous decision-making.

2.1. Composition of multimodal sensors

Robotic multimodal sensors can be categorized into environment-sensing sensors and shape-sensing sensors, including tactile, auditory, pressure, vision, and chemical sensors. These sensing mechanisms primarily rely on electrical, magnetic, and optical properties.

Electrical sensors detect changes in resistance, capacitance, or inductance to perceive the robot's shape or movement. They are cost-effective and easy to integrate but are susceptible to environmental interference that can distort signals. Magnetic sensors use permanent magnets or coils to create magnetic fields for tracking shape, allowing for miniaturization but facing accuracy issues due to interference from other magnetic materials.

Optical sensors utilize optical fiber and Fiber Bragg Grating (FBG) technology, offering flexibility, non-contact capabilities, and biocompatibility, making them ideal for medical applications. However, they tend to be more expensive due to material and equipment costs. Overall, recent advancements have led to integrated sensors that combine multiple principles, enhancing precision and accuracy in robotic applications.

In summary, robotic multimodal sensors incorporate electrical, magnetic, and optical properties to perceive the robot's shape or the movement, position, and direction of objects. Recent advancements have led to the emergence of integrated sensors that combine multiple principles, thereby enhancing precision and accuracy.

2.2. Principles of multimodal sensing systems

The core principle of multimodal sensors is to integrate two or more distinct single-modal sensing modules, enabling multi-dimensional and complementary information acquisition through data collection, preprocessing, and fusion. This integration mitigates limitations of single-modal perception, like illumination interference and signal noise, enhancing accuracy and completeness.

The specific workflow consists of three key steps. First, multi-source data collection occurs as different sensor modules, like cameras and pressure sensors, gather data synchronously or asynchronously. For example, agricultural robots collect crop spectral data with hyperspectral cameras alongside soil data from humidity sensors. Second, data preprocessing standardizes heterogeneous data by filtering noise, unifying dimensions, and aligning time and space for accurate correspondence. Finally, multimodal data fusion integrates the preprocessed data into unified information using advanced algorithms. Common fusion methods include feature-level fusion for

DOI: 10.54254/2755-2721/2025.LD28956

high-precision scenarios and decision-level fusion, which combines independent decision results from each modality, enhancing reliability in safety-sensitive applications [1].

3. Application of multimodal perception systems in autonomous decision-making

Currently, due to the excellent characteristics of multimodal perception systems, robots in many fields have chosen multimodal sensors as their perception hubs. The following are three types of latest multimodal perception robots.

3.1. Bionic soft robots

The human hand is highly intelligent in recognizing and handling objects of different sizes and shapes. Recent advancements in bionic soft robots have led to the development of an intelligent system that integrates capacitive and triboelectric sensors, enabling autonomous operation and multimodal perception. This system employs distributed sensors to perceive and memorize multimodal information, facilitating robot positioning and adaptive grasping. In this process, the multimodal perception system can sensitively capture information and fuse it at the feature level, thereby realizing cross-modal object recognition and highly enhancing recognition capabilities [2].

Meanwhile, to achieve appropriate positioning of randomly distributed objects, integrating a guidance system is inevitable. However, traditional guidance systems based on cameras or optical sensors have limited environmental adaptability, high data complexity, and low cost-effectiveness. Now, by integrating ultrasonic sensors with flexible electrical sensors, a soft robotic perception system with long-range target positioning and multimodal cognitive capabilities has been developed. Ultrasonic sensors can detect the shape and distance of objects through reflected ultrasonic waves. Thus, the robotic manipulator can be positioned to appropriate positions for object grasping. During this period, ultrasonic and triboelectric sensors can capture multimodal sensory information such as the top contour, size, shape, hardness, and material of the object. Finally, these multimodal data are fused for deep learning analysis, thereby greatly improving the accuracy of object recognition [3].

As shown in Figure 1 (a), it is a four-finger soft gripper developed by Suzumori et al. from Yokohama National University, Japan. Each finger, with a diameter of 12mm, consists of 3 pneumatic chambers and multiple sensors of different modalities, and can bend in any direction, enabling dexterous grasping of objects of various shapes (such as beakers filled with liquid, metal wrenches, etc.). It can even tighten small bolts without other additional tools. As shown in Figure 1 (b), it is a six-claw soft gripper composed of a 3-layer soft structure made by Ilievski et al. from Harvard University using embedded pneumatic network technology. By sensing multimodal data, the gripper can control air pressure to achieve forward and reverse bending, thereby adapting to the surface curvature of the grasped object. As shown in Figure 1 (c), it is a variable effective length soft gripper designed by Hao Yufei et al. from Beihang University, which can adjust the effective length of the gripper using nylon ropes to grasp objects of different shapes. Figure 1 (d) shows a modified version based on the characteristics of octopus tentacles, which changes the force according to the data from friction sensors and acoustic sensors at each port to realize the adsorption and twine actions of octopus hands. It can be seen that multimodal perception systems are widely used in soft robots of different forms [4].

DOI: 10.54254/2755-2721/2025.LD28956

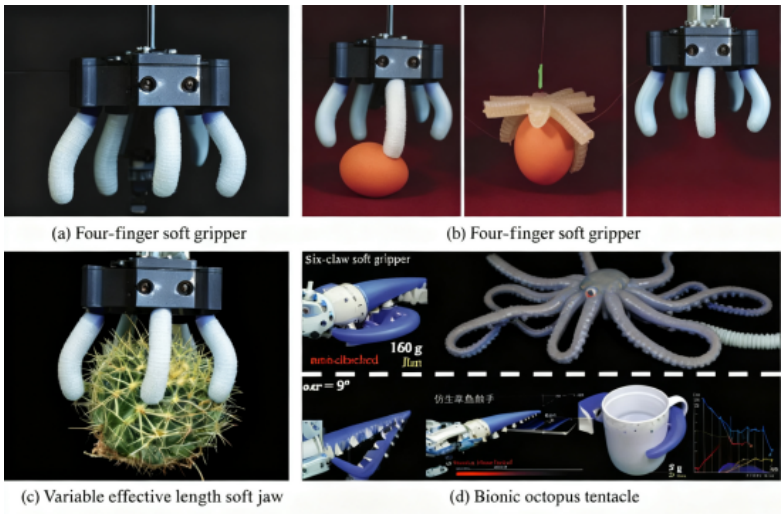

Figure 1. Soft gripper example [4]

3.2. Intelligent agricultural robots

Agricultural robots are typical representatives of new productive forces. By integrating automation and intelligent technologies, they provide strong support for alleviating the structural shortage of traditional agricultural labor, improving production efficiency, and reducing costs. However, most agricultural robots in the past lacked adaptability to complex agricultural environments and still had limitations in facing variable, uncertain, and unstructured agricultural scenarios. Now, multimodal fusion perception technology is used as the "perception hub" of intelligent agricultural robots. Through spatiotemporal collaborative perception and multi-source information fusion of heterogeneous sensor arrays, it provides support for reliable decision-making and successful action execution. The real-time performance and accuracy of multimodal fusion perception technology directly determine the reliability of autonomous decision-making and the accuracy of action execution of agricultural robots.

In the application scenario of unmanned combine harvesters, object recognition and classification involve real-time identification of wheat maturity and lodging areas in the field. When lidar detects lodging wheat ears, the system automatically adjusts the header height and drum speed, reducing grain loss compared with traditional harvesting methods; a multi-source positioning system integrating Global Navigation Satellite System (GNSS) with real-time kinematic positioning technology, visual odometry, and high-precision inertial navigation realizes stable navigation of equipment. In areas where GNSS signals are lost, lidar matching technology and pressure sensors monitor the position and attitude of the machine body, enabling the harvester to maintain low-error row-following accuracy even in muddy fields [5].

In the application scenario of robots autonomously identifying tomato main stems, robots often need a large amount of image data to identify tomato main stems. Moreover, to strengthen the difference between tomato plant main stems and the background, in addition to obtaining RGB images based on a broad visible light band, it is necessary to additionally obtain images of specific strong reflection bands of the stems. The multimodal perception system plays a key role here: first, collect multimodal image data and perform weight assignment preprocessing on it; then, fuse the multimodal features in the images; finally, output the fused images to the recognition and decision module. Figure 2 shows the acquisition process of multimodal images: first, use spectrum to identify the specific spectrum of the main stem; then, obtain different types of images with RGB cameras

DOI: 10.54254/2755-2721/2025.LD28956

and NTIR cameras respectively; finally, fuse the three features to obtain image data with multimodal features. The images processed by multimodal technology have obvious main stem features, which can significantly improve the efficiency and accuracy of agricultural robots in work [6].

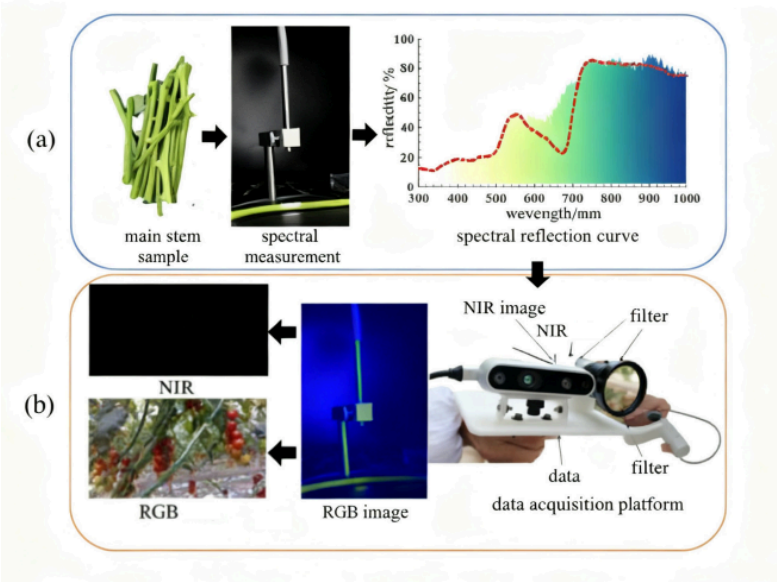

Figure 2. Acquisition of multimodal images [6]

3.3. Medical and health robots

Minimally invasive surgical robots have attracted widespread attention due to their many advantages. However, the lack of sufficient perception capabilities remains a major problem for robots. In this work, a multimodal perception system composed of ultrasonic sensors, capacitive sensors, and triboelectric sensors not only achieves the minimum sensor size but also develops a wireless vibration feedback wristband and a digital twin interface to provide multimodal feedback without interfering with operations. Experimental results show that the perception feedback system improves the safety of surgical robots.

The sensing module can also be applied to the preliminary detection of subcutaneous abnormal tissues. The recognition accuracy based on ultrasonic echoes and convolutional neural networks is 91.6%, which can be used as a preliminary diagnostic reference [7]. In addition to detecting subcutaneous abnormalities, skin disease diagnosis based on multimodal data feature fusion is also an emerging application. First, the extracted lesion size information is used as a feature of text data; then, feature extraction is performed on image and text data respectively. Swin Transformer is selected as the core algorithm for image feature extraction, and a multi-layer perceptron (MLP) is used to process text features. Finally, the extracted image features and text features are fused through a fully connected layer (FC), and the generated global information is used for skin disease classification. The results show that such a classification model significantly improves the efficiency and accuracy of lesion classification [8].

In addition, some scholars have proposed a self-powered electronic skin based on mechanical metamaterials with multimodal fusion perception capabilities and shape memory reconfigurability. The e-skin realizes bionic nonlinear mechanical behavior and mechanical reconfigurability imitating target human tissues. Its integrated perovskite-based elastic sensors can achieve high-precision acquisition of physiological movement, auditory, tactile, and pre-contact distance signals.

DOI: 10.54254/2755-2721/2025.LD28956

Furthermore, by imitating the integration and interaction functions in biological multi-sensory neural networks, the system also realizes advanced cognitive functions of acquiring, recognizing, and integrating cross-modal information. It demonstrates the application of electronic skin in motion monitoring, speech recognition, and somatosensory game operations [9].

In summary, it can be seen that multimodal perception systems have many applications in many robotic fields. Their development will greatly improve the intelligence level of robots and have great research value.

4. Current challenges and development directions

Currently, the first core challenge of multimodal perception systems is the existence of a gap between heterogeneous modalities, resulting in high costs for aligning modalities of different dimensions.

The heterogeneity of multimodal data is not only reflected in differences in physical forms (e.g., two-dimensional matrices of images and one-dimensional pressure signals of touch) but also in the separation of semantic spaces. For example, "edge features" in visual data and "frequency features" in voiceprint data are difficult to align directly, leading to the failure of traditional feature splicing methods in complex scenarios. Although Microsoft's BABEL framework improves the recognition accuracy of six types of sensor data (Wi-Fi, millimeter waves, etc.) by 22% through scalable modal alignment technology, its design relying on pre-trained modal towers still faces the problem of cross-modal semantic drift, with high time and technical costs [10].

Second, there is a contradiction between real-time performance and computing power. For example, in medical scenarios, if a multimodal perception system is used to operate robots to complete remote surgery, real-time multimodal image fusion during surgery needs to control the delay within 50ms, but existing GPU clusters are difficult to meet this demand.

Future advancements must focus on developing sensing materials with good mechanical compliance and physicochemical properties, creating uniform high-precision multimodal sensors, and achieving large-area integration of sensor arrays. Addressing these issues will enhance the speed and accuracy of information acquisition and transmission, thereby alleviating current challenges.

5. Conclusion

The ability of multimodal sensing systems to perceive various external elements is crucial in robotic applications. This paper mainly introduces the composition, principles, and current emerging applications of multimodal sensors, and also proposes current main core challenges and development directions. However, the field of robotics is a rapidly developing forward-looking field. With the development of new motion structures and motion principles, the demand for sensing shapes and environmental interactions is increasing, which promotes research on new sensing methods and new sensor designs. In the future, further focus should be placed on innovation in sensing methods and optimization of sensor design. The above challenges of multimodal sensors still require joint efforts from scientific researchers and industrial engineers.

However, this paper still has shortcomings, such as not in-depth discussion on the adaptability and optimization schemes of perception modules in extreme environments (e.g., high temperature, strong electromagnetic interference), insufficient analysis of lightweight design of multimodal fusion algorithms, and inadequate analysis of the application of multimodal systems in various fields.

DOI: 10.54254/2755-2721/2025.LD28956

References

[1] Mao, J., Nie, X., & Zhu, H. (2025). Research and progress of multimodal sensors in robotics. Science and Innovation, (12), 29-32. https: //doi.org/10.15913/j.cnki.kjycx.2025
[2] Wang, T., Jin, T., Lin, W., et al. (2024). Multimodal sensors enabled autonomous soft robotic system with self-adaptive manipulation. ACS Nano.
[3] Qiongfeng, S., Zhongda, S., Xianhao, L., et al. (2023). Soft robotic perception system with ultrasonic auto-positioning and multimodal sensory intelligence. ACS Nano.
[4] Liu, W., Wang, Y., Duo, Y., et al. (2024). Research progress on interaction of soft robots based on flexible sensing. Robotics.
[5] Wei, P., Cao, S., Liu, J., et al. (2025). Embodied intelligent agricultural robots: Key technologies, application analysis, challenges, and prospects. Smart Agriculture (Chinese and English).
[6] Liu, C. (2023). Research on the recognition and obstacle avoidance tracking method for tomato main stem aimed at autonomous operation of robots [Master's thesis, Sichuan Agricultural University].
[7] Li, D., Ji, T., Sun, Y., et al. (2025). A full-range proximity-tactile sensor based on multimodal perception fusion for minimally invasive surgical robots. Advanced Science (Weinheim, Baden-Württemberg, Germany).
[8] Zeng, Z. (2025). Research and implementation of skin disease diagnosis based on multimodal data feature fusion [Master's thesis, Sichuan Normal University].
[9] Li, N., et al. (2024). Metamaterial-based electronic skin with conformality and multisensory integration. Advanced Functional Materials.
[10] Hong, C., Zhiquan, F., Jinglan, T., et al. (2023). MAG: A smart gloves system based on multimodal fusion perception. CCF Transactions on Pervasive Computing and Interaction.