Soft Robotic Grasping Perception System Based on Multimodal Intelligent Sensor Fusion

Haoran Yang

{235779@stu.hebut.edu.cn}

School of Mechanical Engineering, Hebei University of Technology, 300401, Tianjin, China

Abstract. Soft robots have become a key technology for precise grasping and human-robot collaboration in complex environments due to their flexible structures and high safety. Multimodal intelligent sensor fusion has shown significant advantages in perception and decision-making of soft robots. Soft robots can achieve high-precision grasping and optimized movements in dynamic, obstructed or uncertain environments by integrating tactile, visual and inertial data from multiple sources and combining it with multimodal intelligent sensing algorithms. At the hardware level, tactile, visual and inertial sensors form the basis of multidimensional perception. At the algorithmic level, fusion algorithms, adaptive fusion strategies, and learning-driven action optimization have enhanced the robustness and adaptability of the system. Multimodal intelligent sensor fusion technology is widely used in scenarios such as industrial precision operations, rehabilitation assistance, and underwater grasping. This article explores the development trends and future research directions of multimodal intelligent integration in soft robots. It provides a reference for the next generation of highly autonomous and adaptable intelligent soft robot systems.

Keywords: Soft Robots, Multimodal Intelligent Sensor Fusion, Perception Systems, Flexible Sensors, Intelligent Fusion Algorithms.

1 Introduction

In recent years, due to their flexible structure, high adaptability and safety, soft robots have attracted much attention. They can conform to the target's complicated shape and resist external disturbance, which can decrease the probability of damage during the contact.

However, to realize the above advantages, the soft robot needs to have a better environment perception. The single sensor mode makes it limited. In a complex environment, the problem of information loss and congestion in the process of exploration is inevitable. For example, it is difficult for tactile sensing to achieve accurate positioning, and inertial information cannot reflect the grasping dynamic process completely. The above problems can be solved by adopting a strategy called multimodal intelligent sensing fusion to improve grasping performance. The soft robot can realize stable grasping and motion optimization in an uncertain environment by combining tactile, visual and inertial sensing with a multimodal intelligent sensing fusion

ICMEEA 2026, April 24-26, Singapore, Singapore

DOI 10.4108/eai.24-4-2026.2364827

algorithm. One hand, the flexible sensor and deformable visual soft membrane can collect multi-source information. On the other hand, the fusion at the algorithm level and dynamic optimization enhance the adaptability of the system. The application of this multimodal, intelligent sensing fusion technology is very broad. It exists in industrial precision operation, assistive rehabilitation devices, underwater operation, etc. It can effectively improve the grasping success rate and environment adaptability.

This article analyses the integration of hardware and fusion of algorithms and typical application of multimodal intelligent perception fusion technology for soft robot systematically. It provides a good reference for the development of high performance and intelligent soft robotic system.

2 Multimodal Intelligent Sensing Hardware Integration

Multimodal intelligent sensing hardware is crucial for the grasping perception capabilities of soft robots. The design of such hardware should focus on two key principles: first, compatibility with the compliance of soft structures; second, meeting the perception requirements throughout the entire grasping task. The sensor needs to maintain its operational flexibility when the soft tooling undergoes deformation. At the same time, it needs to integrate tactile, visual and inertial information to overcome the limitations of a single sensing method in precise operations, providing a stable and multi-dimensional data foundation for the formulation of grasping strategies.

2.1 Tactile Sensing

In soft robotics grasping systems, tactile sensing is the primary method of perception. It can collect information on the pressure distribution and deformation of flexible structures. This enables precise grasping and optimized motion. Because fixed structures are flexible, tactile sensors must be highly flexible and precise. They must also fuse multiple modalities. The current approaches mainly involve capacitive and triboelectric sensors. Capacitive sensors use flexible electrodes and microstructure dielectric layers to detect changes in external forces. The soft humanoid hand developed by Weiner et al., for example, integrates capacitive pressure, 3D Hall shear force, proximity sensing and accelerometer units. It used only 35.64% of the available energy when grasping 31 objects every day and did not damage any fragile items [1]. This approach greatly surpasses conventional tactile approaches. Unlike these tactile sensors, those can always provide contact information for the following multimodal fusion and grasping decision. Triboelectric sensor adopts the principle of triboelectric charging and electrostatic induction to produce signals, so there is no need for power supply. It has advantages of low power consumption, high sensitivity and strong reliability. Its flexible structures and mechanical transmission mechanism convert the gripper bending into electric signals, which realize the dynamic motion monitoring and the closed-loop control. The piezoresistive bending sensor proposed by Shi et al. shows good signal stability in the simulated industrial sorting application. Its output amplitude is linearly related to the bending angle, which offers information for the motion control [2]. Usually, soft robotic gripping systems will integrate the information from more than one type. The soft gripper can output information of contact force, object surface and relative motion at the contact position in real time. This integrated system improves the grasping stability and offers more information for the following algorithm.

2.2 Vision Sensing

Visual sensing is another way to help soft robots grab objects. This method can provide information of spatial position and contours of target objects as well as the motion states of the target objects. After the contact phase started, the visual obstruction became a critical issue. In recent years, selective light-transmitting materials of soft membrane have been a good solution to solve this problem. For instance, Yin et al. fabricated a soft Ecoflex film with only 0.01 millimeters in thickness. This film contains ultraviolet-excited phosphorescent particles and infrared-transmitting dyes in. This 1 mm-thick Ecoflex film can transmit 860 nm infrared light with a transmittance of 91%. This soft material achieves the compromise between physical elasticity and sensing functionality by embedding ultraviolet phosphorescent particles and infrared-transmitting dyes in the material. Meanwhile, this material can transmit infrared light and block visible light. When a vision system integrated with ToF (time-of-flight) camera and optical markers are used, the positioning error of contact points is at least 1 millimeter when the distance is 50 millimeters. When the distance reaches 100 millimeters, the error becomes 31 millimeters [3]. Vision systems based on time-of-flight (ToF) sensors with selective light-transmitting materials can be used to sense continuously in the whole process. In addition, these systems can also provide data support for other data from different modalities in intelligent perception fusion. It enhances the environmental adaptability and operational reliability of soft robots completing complex manipulation tasks.

2.3 Inertial Sensing

The inertial sensor collects data about the dynamic state of the actuator. It is an effective complement to motion monitoring based on touch and vision. When visual and tactile perception are limited, the inertial measurement unit (IMU) offers stationary input for grasping control. Lin et al. connected modular soft actuators (MSA) with an inertial measurement unit (IMU) to build a dynamic mapping model. The gripper's opening range and target dimensions were adaptively matched. It solved the technical problem of measuring the angle of the flexible joint [4]. Martin et al. proposed to approximate the soft continuum actuator as a rigid link series structure. They placed inertial measurement units (IMU) on the actuator to collect attitude data. They used spherical linear interpolation and a piecewise constant curvature model to collect attitude data. They used this data to accurately reconstruct the curved state. In the experiment using the 660-millimeter soft robotic arm, the end-effector position estimation error did not exceed 10% of the arm length. It proved the effectiveness of the method in identifying dynamic states [5]. Inertial data can be combined with visual and tactile perception. The combination of these three modalities improves the temporal consistency of the system during grasping actions and improves the accuracy of the state estimation. Therefore, habitual perception is the most important "body perception layer" in soft robotics. In addition, it is also very important to temporal compensation and state prediction in multimodal intelligent fusion. It provides the necessary support for stable operation in the working environment.

3 Modal Sensing Fusion Algorithm

Multimodal intelligent fusion algorithms improve the adaptability and robustness of soft robots. These algorithms are based on information from multiple perceptual sources, such as tactile, visual and inertial information. Compared with single sensor mode, these algorithms break through the application limitation in occluded and dynamic environment. Soft robots' intelligent fusion algorithm utilizes the information from multiple modalities. Soft robots grasp objects accurately in the occluded environment.

3.1 Feature Fusion Algorithm

Feature fusion algorithm is the key component of multimodal perception. It fuses visual data, depth and position information into one feature space, which enables the robot to have a correct perception and take right actions in the environment. Now there are two ways to fuse features: early fusion and deep shared representations. Early fusion means to integrate features in preprocessing stage. For example, Nan et al. integrated bird`s view (BEV) image feature and positional information extracted from 3D LiDAR into one feature by two-dimensional convolutional layer and fully connected layer. Then optimize and integrate them into one vector. It can quickly fuse the information. Deep shared representation means to use deep networks to promote interaction between modalities. This study uses recurrent neural units (GRUs) to model temporal dependencies and capture dynamic environmental changes. Experiments have demonstrated an improvement in navigation success rates of over 30% compared to the single-modal TD3 algorithm [6]. Feature fusion algorithms are an effective way of integrating multimodal information. However, their computational complexity and insufficient real-time performance remain a constraint. Future improvements can be achieved by designing lightweight networks and optimizing them collaboratively using adaptive fusion strategies.

3.2 Adaptive Fusion Strategy

The adaptive fusion strategy aims to automatically adjust the fusion weights of each modality based on the reliability, confidence level and dynamic changes in the environment. Thereby enabling on-demand integration of information. Compared to traditional fixed-weight methods, this strategy maintains system robustness and stability even when sensors are blocked or data is missing. Typical methods include attention-based fusion and confidence-based fusion. Kang et al. developed a vision-servo control system for tendon-driven soft robots. They used a three-dimensional force sensor to capture contact force signals, as well as a monocular camera to extract image features. These allowed them to construct a contact disturbance estimation model. This method can assess the reliability of the visual modality using contact force signals and can adjust the force compensation weights accordingly. They also used Chebyshev polynomials to simplify high-dimensional modelling and incorporated the Slottin-Lee adaptive law to speed up parameter convergence. This validated the effectiveness of the method across various complex contact scenarios [7]. This approach offers greater flexibility when processing redundant information from multiple sources and executing real-time tasks. Adaptive fusion strategies can effectively improve the anti-interference capabilities and environmental adaptability of multimodal systems. However, convergence speed and computational overhead remain major challenges in high-dimensional feature spaces.

3.3 Learning-Driven Action Optimization

Learning-driven action optimization is represented by deep reinforcement learning and imitation learning. This aims to achieve autonomous optimization and dynamic adjustment of grasping strategies based on multimodal perception features. This class of methods uses the fused perception results as state inputs. By learning action decisions through the policy network, it achieves closed-loop optimization from “perception-decision-execution.” Jia et al. propose integrating multimodal perception fusion with deep reinforcement learning. This approach optimizes closed-loop control in mobile robot navigation research. [8]. This method uses two-dimensional LiDAR to acquire environmental depth information and a color camera to provide visual features. The LiDAR data is calibrated through coordinate transformation and YOLOv8 detection results to form a fused perception input. It is fed into a policy network based on the soft actor-critic (SAC) algorithm, which produces velocity and steering control commands. It achieves stable navigation in dynamic obstacle scenarios by employing a multidimensional reward function that combines obstacle penalties with target distance rewards. This validates the effectiveness of a closed-loop reinforcement learning optimization driven by multimodal perception. Soft robots can be endowed with adaptability and autonomous learning capabilities through learning-driven control optimization. This method suffers from significant sample dependency and high training costs. Future research should focus on developing more efficient strategies.

3.4 Learning-Driven Action Optimization

This section provides an overview of different intelligent fusion algorithms. As shown in Table 1. It includes fusion algorithms, adaptive fusion strategies, and learning-driven control optimization. These form the core of the algorithmic framework for multimodal intelligence fusion in soft robotics.

Table 1. Core Elements Comparison of Multimodal Intelligent Sensing Fusion Algorithms

Algorithm	Core Objective	Key Technologies	Core Challenges
Feature Fusion Algorithm	Map heterogeneous data to a unified feature space	2D Convolution, Layer Normalization, GRU	High computational complexity and insufficient real-time performance
Adaptive Fusion Strategy	Automatically adjust the fusion weight of each modality based on modal reliability/confidence	Contact disturbance estimation model, Chebyshev polynomials	Slow convergence and high computational overhead in high-dimensional feature spaces
Learning-Driven Action Optimization	Realize autonomous action optimization and the "perception-decision-execution" closed loop	2D LiDAR + Color Camera, YOLOv8, SAC Algorithm, Multi-dimensional Reward Function	Strong sample dependency and high training costs

4 Typical Application Scenarios

4.1 Industrial Precision Operations

In industrial manufacturing and assembly, precision operations demand more stability from soft robots. Because traditional actuators, which are fixed configuration, must be replaced or reprogrammed, this results in long downtime and a large cost. Therefore, they are not ideal for small-batch, high-mix operations. Abhishek et al. designed a modular, reconfigurable actuator to provide an adaptable solution for industrial applications. This device consists of a base and two active tools. This device utilizes a four-bar linkage mechanism and pneumatic drive system. This configuration can be switched to workpieces of different sizes and shapes within less than a minute without reprogramming. This device can work with industrial robots like the ABB IRB1410 and greatly reduces the cost of commissioning. Their efficiency and versatility have been demonstrated in sorting and assembly operations in small-batch manufacturing [9]. However, this actuator does not include any sensing units and only works on pre-programmed sequences. When there are variations in workpiece orientation or surface characteristics, positioning errors and grasping failure will occur, which reduces the adaptability of this actuator in different working scenarios. This issue can be solved by a multimodal intelligent sensing fusion system. By embedding visual, tactile, and inertial sensing modules inside this actuator, the intelligent sensing fusion system can achieve the synergistic fusion of three modes of vision, touch, and inertial information. This overcomes the shortcomings of conventional actuators and enables soft robots to integrate both adaptable configurations and self-adaptability in industrial precision operations, which greatly enhances the operation precision and system reliability.

4.2 Rehabilitation-Assisted Grasping

In rehabilitation, assisted grasping should be safe and patient averse. Multimodal fusion can effectively correlate rehabilitation demands for patients by overcoming sensory impairment and poor motor coordination. As shown in [10], rehabilitation assisted by mechanical structures is feasible. Because of the soft properties of robotic materials, the speed of the opening and closing of the gripper is coordinated with the range of motion of the patient's joint, thereby avoiding the discomfort generated by large joint movement. Feature-level fusion refers to the fusion of tactile pressure and inertial posture information based on the temporal alignment method to ensure the perception consistency of information. Imitation learning refers to the simulation of healthy grasping motion to assist patients in forming the logic of grasping according to their own ability. Reinforcement learning adjusts the action parameters iteratively to improve the accuracy of grasping by patients continuously. Multimodal fusion takes multidimensional information to form a safety barrier. It empowers patients to regain autonomy in grasping through adaptive algorithms, which also meets the fundamental needs of healthcare rehabilitation.

4.3 Underwater Environment Capture

The low light intensity and interference of water flow led to positioning errors of vision and interference of touch sensing in underwater environment. The core of solving above bottlenecks of underwater sensing is multimodal fusion. That is, vision uses translucent and flexible membrane and ToF camera filtering algorithm to reduce positioning errors of turbidity, to know the exact position of underwater targets, while tactile sensing uses high sensitivity capacitive array to sense the softness of biological objects and hardness of samples to avoid damage during grabbing. Yang Wenna et al. advanced an equivalent Coulomb charge distribution calculation method to solve this. That is, changing the non-Coulombic potential generated by underwater electromagnetic interference into equivalent Coulomb charge distribution, and then using Gauss's law to calculate the strength of interference field; based on this, the design of tactile sensor shielding structure can be optimized to counteract the interference of electromagnetic interference [11]. Selective fusion strategies reduce the positioning errors and interference compensation weights. Reinforcement learning focuses on the dynamics of water flow and electromagnetic interference. The integration of multimodal fusion and field theory computational methods overcomes the traditional problems of underwater perception and enhances the anti-interference of sensors, which provides reliable technical support for the development of marine exploration and underwater biological surveys.

5 Conclusion

The flexibility of soft robots allows them to be used for grasping in various environments and collaborative robotics with humans. Multimodal intelligent sensor fusion is an effective approach to overcome the limitation of single sensor technology. It also can provide high-precision perception, adaptive decision and stable action. This paper proposes the technical framework of multimodal fusion in a comprehensive way. In the aspect of hardware, tactile, vision and inertial sensors provide an integrated perception support. It can fulfill the requirement of contact force monitoring, spatial localization, motion tracking and environment adaptation. At the level of algorithm, feature-level fusion unifies the data representation and selective fusion balance the accuracy and efficiency. Algorithm establishes a ‘perception-action’ closed loop and improves the grasping robustness in dynamic environment.

There are still many challenges existing in the field of multimodal intelligent sensing fusion, such as insufficient hardware robustness, limited real-time algorithm performance and high cost. With the rapid development of durable material and lightweighting algorithms, the next generation of multimodal soft robot can adapt to work in extreme environment. It will be widely used in industrial, medical and underwater applications. Based on that, the technological foundation of highly autonomous, adaptable and intelligent robotic system will be established.

References

[1]. Weiner, P., Hundhausen, F., Grimm, R., Asfour, T.: Detecting grasp phases and adaptation of object–hand interaction forces of a soft humanoid hand based on tactile feedback. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3979–3986. IEEE, Prague, Czech Republic (2021)

[2]. Shi, Q., Sun, Z., Le, X., Xie, J., Lee, C.: Intelligent soft robotic gripper enabled by multimodal sensors and deep learning. In: 2023 IEEE International Conference on Nano/Micro Engineered and Molecular Systems (NEMS), pp. 19–22. IEEE, Jeju, Korea (2023)

[3]. Yin, J., Campbell, G.M., Pikul, J., Yim, M.: Multimodal proximity and visuotactile sensing with a selectively transmissive soft membrane. In: 2022 IEEE International Conference on Soft Robotics (RoboSoft), pp. 802–808. IEEE, Edinburgh, UK (2022)

[4]. Lin, S.-G., Chang, H.-T., Chang, J.-Y.: IMU-based robotic finger force sensing with modular soft actuators. In: 2024 IEEE International Magnetic Conference – Short Papers (INTERMAG Short Papers), pp. –. IEEE, Rio de Janeiro, Brazil (2024)

[5]. Martin, Y.J., Bruder, D., Wood, R.J.: A proprioceptive method for soft robots using inertial measurement units. In: Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 9982185. IEEE, Kyoto, Japan (2022)
[6]. Nan, Z., Nam, H.: Multimodal feature fusion for deep reinforcement learning-based mobile robot navigation. In: 2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. –. IEEE, Ansan, Korea (2025)
[7]. Kang, X., Xu, F., Han, L., Wang, H.: Adaptive visual servo of soft robot with interaction estimation and compensation. IEEE Transactions on Automation Science and Engineering 22, 18395–18404 (2025)
[8]. Jia, T., Deng, H., Zhan, J., Zhang, L.: A deep reinforcement learning-based robot navigation method using sensor information fusion. In: 2025 40th Youth Academic Annual Conference of Chinese Association of Automation (YAC), pp. 559–564. IEEE, Zhengzhou, China (2025)
[9]. Abhishek, M., Jaman Sahaul, N., Vishnuvardhan, S., Rochin Antony, S., Ram Kumar, A., Sivaguru, J.: Design, development and deployment of reconfigurable end-of-arm gripper for industrial robots. In: 2023 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), pp. –. IEEE, Coimbatore, India (2023)
[10]. Wang, C.-J., Deng, H.-L.: Application of soft robots in rehabilitation training. In: Proc. International Conference on Robotics, Automation and Medical Rehabilitation (ICRAMR), pp. 1–7. IEEE, Huainan, China (2024)
[11]. Yang, W.-n., Xiong, Z.-h.: Calculation method of equivalent Coulomb charge distribution and its application. Journal of Beijing University of Technology 49(4), 1–6 (2023)