Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/154/2025.TJ23114

Vertical Applications of Multi-modal Sensor Fusion: A Comparative Study of Medical Robots and Industrial Robots

Jiaxing He

Guangdong University of Technology, Guangzhou, China
16626401302@163.com

Abstract: Multimodal sensor fusion technology significantly improves the perception and decision-making capabilities of medical and industrial robots by integrating multi-source information such as vision, touch, and mechanics. This article conducts a comparative analysis of the divergent applications and shared characteristics within two predominant domains: medical robotics, which emphasizes surgical safety and employs tactile-visual collaborative technologies to enhance operational precision while addressing data compliance challenges through a privacy protection framework; and industrial robotics, which prioritizes efficiency and safety in dynamic environments by integrating dynamic vision systems and high-precision ranging devices to facilitate real-time obstacle avoidance and fault diagnosis. The study found that although the scene goals are different (medical focuses on biocompatibility, industry focuses on cost efficiency), both face challenges such as sensor redundancy, data heterogeneity, and long-term stability. In the future, it is necessary to promote cross-domain technology interoperability, develop a lightweight adaptive fusion framework, and build an ethical and standardization system to promote the universalization and large-scale application of multimodal fusion technology.

Keywords: Multi-modal sensor fusion, medical robot, industrial robots, human-robot collaboration

1. Introduction

Multimodal sensor fusion has emerged as a pivotal research area in medical and industrial robotics. In medical applications, precise manipulation and real-time feedback are paramount for patient safety. Industrial settings increasingly demand efficient collaboration and human-computer interaction in complex environments. Traditional unimodal sensors struggle to meet the requisite precision and robustness due to their limited perceptual dimensionality. Multimodal fusion compensates for these limitations by integrating multi-source data—vision, tactile, and mechanical inputs—and fosters cross-domain technological innovation.

Drawing upon medical robotics, specifically the Da Vinci system, this study affirms the incorporation of tactile and visual modalities. However, hardware limitations and data acquisition issues impede research progress. Conversely, industrial robots employ multi-modal collaboration—dynamic vision and lidar—to achieve environmental perception and real-time decision-making in intricate tasks. Despite disparate applications, both domains grapple with sensor redundancy, data heterogeneity, and privacy. This article contrasts technical approaches, applications,

DOI: 10.54254/2755-2721/154/2025.TJ23114

and challenges in multi-modal sensor fusion across medical and industrial sectors, exploring shared needs and divergent trends to inform cross-domain technological innovation.

2. Application of multimodal sensor fusion in medical surgery

2.1. Fusion of haptics and vision in surgical robots

In the realm of visual perception, the Da Vinci Xi system facilitates three-dimensional depth perception via stereoscopic vision, thereby enhancing surgical trajectory planning [1]. However, its clinical orientation makes it difficult for research institutions to obtain hardware and data, limiting research progress. In pursuit of this objective, the Da Vinci Research Kit (dVRK) was developed, enabling researchers to comprehensively engage with the control hierarchy and data flow. This initiative offers an open-source software and hardware platform for international research collectives, thereby reducing barriers to research and fostering collaborative efforts.[2]. Therefore, multi-tool switching and automated tumor resection technology are realized. By designing interchangeable instrument interfaces (such as Clevis Mount and Jaw Mount), the system can directly complete the autonomous switching of palpation probes, scalpels, clamps, and syringes in the body cavity without repeatedly entering and exiting the cannula. The tool change adapter (TCA) facilitates the robotic arm's execution of 30 consecutive grasping actions devoid of visual feedback, utilizing passive guide grooves and self-locking mechanisms. This innovative design markedly diminishes the temporal inefficiencies associated with conventional tool interchange, thereby enhancing surgical efficacy [3].

Real-time intraoperative fluorescence (e.g., with indocyanine green) guidance has shown great potential in helping guide surgeons in both simple and complex surgical interventions, and as a result, fluorescence imaging was introduced to the Da Vinci Surgical System in 2011 and has been standard equipment since 2014[4].

Within the realm of tactile sensors, fiber Bragg grating (FBG) exhibits distinct advantages, including its compact dimensions, excellent biocompatibility, elevated sensitivity, capacity for sterilization, cost-effectiveness, and diminutive size [5]. It solves the problem of insufficient force measurement in minimally invasive robotic surgery [6]. However, temperature sensitivity and dynamic response are insufficient, and future research will explore the potential of force feedback in performance improvement, surgical training, and skill assessment. Piezoelectric sensors have fast dynamic responses and are suitable for vibration signal detection, such as suture tension monitoring. Capacitive sensors are highly sensitive and can be used in tactile skin development. However, piezoelectric and capacitive sensors experience stress relaxation under long-term pressure, which may affect the accuracy of measurements. A comprehensive assessment of sensor performance across various in vitro and in vivo evaluations, the advancement of algorithms for stress relaxation compensation, enhancements in sensor sealing technologies, and a thorough examination of their efficacy and safety in real-world clinical applications are imperative [7].

Robot-assisted suturing cases show that in tissue puncture tasks, direct force feedback can significantly reduce the maximum force applied from 2.54 N to 2.49 N, reduce tissue damage and the number of impacts from 1.12 times to 1.08 times, and shorten the task completion time from 57.05 s to 51.73s. In knotting tasks, visual force feedback helps to improve the quality of suture knots and the consistency of force, but when used alone, it is easy to cause suture breakage or too loose suture knots. Optimal outcomes are attained when direct haptic feedback is integrated with visual feedback mechanisms (evidenced by only 3 out of 75 ligation tasks exhibiting looseness and no instances of breakage). This dual feedback approach not only mitigates tissue trauma and adverse effects through direct haptic input but also enhances the stability of suture knots and maintains consistent force application, facilitated by visual feedback [8].

DOI: 10.54254/2755-2721/154/2025.TJ23114

Surgical robots are designed to improve surgical precision and flexibility. Surgeons use multimodal sensing technology to control robotic arms. Compared with single-mode sensors, optical and mechanical dual-mode sensors can enhance the accuracy, safety, and robustness of human-machine interaction systems. The latest sensing technology achieves optical sensing and shadow recognition, mechanical sensing, and touch force detection by building a parallel structure of perovskite and graphene. The integration of these two elements establishes a synergistic framework for instantaneous, multi-faceted feedback, thereby enhancing the dynamics of human-machine interaction [9].

2.2. Application of sensor fusion in medical monitoring

With the popularization of medical IoT devices, the integration of multimodal sensor data (such as physiological signals, environmental parameters, biochemical indicators, etc.) has become the key to improving the accuracy and real-time performance of patient monitoring. The multi-sensor fusion wearable health monitoring system allows the execution of biometric and medical monitoring applications. It offers haptic feedback modalities and intuitive visual indicators contingent upon the user's health metrics. The aggregated biometric data can be utilized for real-time monitoring of the patient's health condition or to acquire critical information for subsequent medical evaluation and analysis [10].

However, traditional centralized processing methods are difficult to meet medical data compliance requirements due to the risk of privacy leakage. The privacy protection framework based on federated learning (such as PHMS-Fed) solves this contradiction through distributed collaborative training. The framework uses adaptive tensor decomposition technology while avoiding raw data transmission, significantly reducing privacy risks. In authentic medical datasets, exemplified by MIMIC-III, the system attained a physiological monitoring precision of 0.9386, a privacy safeguarding metric of 0.9845, and a fusion accuracy of 0.9591, surpassing conventional methodologies by 23% to 25% [11].

3. Multimodal sensor fusion in industrial robots

3.1. Sensor types and scenario requirements

Industrial robots achieve accurate perception and efficient collaboration in complex environments through the collaborative work of multimodal sensors.

Within the realm of visual sensing technologies, RGB-D cameras, exemplified by devices like the Kinect, are employed for three-dimensional environmental modeling, obstacle recognition, and human tracking. These systems facilitate the real-time modulation of a robot's motion trajectory, thereby enabling the avoidance of collisions with humans or other entities. Event cameras, such as the dynamic and active-pixel vision sensor (DAVIS), have low latency and high dynamic range for high-speed robotic operations. Combining traditional global shutter cameras and event-based sensors in the same pixel array can significantly improve the response time for real-time scene analysis, especially for dynamic industrial environments [12].

As a distance sensor, lidar is widely used in human perception and is often used in combination with visual sensors. For example, lidar is combined with visual sensors for human detection and behavior understanding to achieve safe and efficient collaboration. Lidar provides high-precision distance measurement and environment modeling capabilities in these applications, helping robots better perceive and understand human behavior [13].

In the realm of industrial wearables, inertial measurement units (IMUs) serve a pivotal role in monitoring operator movements, which are subsequently analyzed and categorized to govern robotic behavior. Notably, IMU sensors are capable of discerning both static and dynamic gestures, which

DOI: 10.54254/2755-2721/154/2025.TJ23114

facilitate the command of the robot, with these movements being processed and classified through artificial neural networks (ANNs) [14]. In addition, Leap Motion is used for gesture recognition in fine-scale collaborative tasks as input commands for the robot system. It can enhance the robot's perception ability and improve the interaction accuracy in combination with Kalman filtering [15].

The core of the design of multimodal sensors for industrial robots lies in scene adaptability and functional complementarity. Technically, RGB-D cameras employ synchronous frame data for 3D modeling of static scenes, offering global geometric insights, though algorithmic optimization is needed to address depth perception noise in reflective, transparent, and distant objects. Event cameras, utilizing asynchronous event streams, excel in microsecond responses to dynamic targets, circumventing motion blur. Lidar, leveraging active laser ranging, provides stable, high-precision point clouds under varied lighting, compensating for RGB-D cameras' limitations in distance and material sensitivity. IMU and Leap Motion, through inertial data, dissect human motion and gesture tracking, establishing a hierarchical interactive framework of intent screening and precise execution.

These sensors, encompassing geometric reconstruction, dynamic response, high-precision ranging, human motion capture, and fine gesture recognition, address the perception requirements of environment-human-robot interaction. The system employs multi-source data fusion—such as lidar+RGB-D's geometric-semantic complementarity and IMU+Leap Motion's action-intention coordination—to enhance robustness.

3.2. Multimodal data fusion methods and case studies

Early fusion (raw data) refers to directly fusing raw data of different modalities into a joint model in the early stage of the model[16].

Late fusion, a decision-layer approach in multimodal data processing, independently trains models on each modality, generating local decisions. Subsequent integration occurs via rules, voting, or weighted methods. This method's flexibility and fault tolerance render it suitable for complex tasks involving heterogeneous sensors or dynamic environments, such as those shared by humans and mobile robots in remotely operated scenarios. The robot's enriched 360-degree view, augmented with interactive elements, directs attention to information-rich areas [17]. Using a 360-degree camera, whose frames are processed using a YOLO-based convolutional neural network (CNN) framework, the perception of human operators and robots is enhanced in some way [18].

Another technique is minimum redundancy-maximum relevance (mRMR). The goal of this method is to find those metrics that minimize data redundancy since removing a feature from a highly interdependent set will not result in a change in the information they provide; at the same time, the method must maximize the relevance to the target class [19]. The algorithm has an unsupervised version (UmRMR) that has been used for predictive maintenance [20] and structural health monitoring of rotating machinery [21].

Mid-term fusion integrates multi-source data via feature concatenation and dimensionality reduction [21]. Principal Component Analysis (PCA), a classic unsupervised method, serves in feature extraction and selection. PCA transforms features, typically reducing dimensionality to retain only those explaining maximal variance, thus compressing and reconstructing the feature space [22]. Employed in industrial contexts like induction motor fault diagnosis, PCA also has nonlinear variants, such as kernel PCA, to enhance nonlinear feature expression [23].

However, with the advent of deep learning, researchers have increasingly favored adaptive fusion frameworks predicated on deep neural networks. For instance, deep learning feature fusion, an adaptive method rooted in DCNN, addresses multi-sensor feature extraction by selecting the appropriate fusion technique relative to the fusion stage. The capacity of DNNs to fuse data across varying layers and stages underpins their adoption in our approach. The adaptive network performs low-level input data fusion, extracts basic features, integrates these into high-level features and

DOI: 10.54254/2755-2721/154/2025.TJ23114

decisions at an intermediate level, and subsequently, at a higher level, recombines features and decisions to yield the final prediction [22].

Despite the promise of camera-radar fusion, its limitations necessitate innovation. I propose a novel proposal-level fusion method for 3D object detection, associating image proposals with radar points in polar coordinates to address coordinate system and spatial attribute disparities. Through iterative cross-attention-based feature fusion layers, adaptive exchange of spatial context information between camera and radar streams achieves robust and focused fusion. Evaluation on the nuScenes test set demonstrates a mean Average Precision (mAP) of 41.1% and a NuScenes Detection Score (NDS) of 52.3%, surpassing the camera-only baseline by 8.7 and 10.8 points, respectively, and outperforming lidar-based approaches [23].

4. Comparative analysis & future development direction

4.1. Comparative analysis

Multimodal sensor fusion in medical and industrial robots exhibits divergent applications and shared imperatives. Medical robots prioritize surgical precision and patient safety, integrating tactile feedback (e.g., fiber Bragg gratings, piezoelectric sensors) with stereo vision for minimally invasive procedures like sutures and biopsies. Technical challenges reside in biocompatible materials and tactile dynamic compensation algorithms, while multimodal health monitoring is constrained by patient data privacy regulations. Industrial robots emphasize collaborative efficiency and dynamic environmental safety, employing event cameras, lidar environmental modeling, and wearable devices for real-time obstacle avoidance and fault diagnosis. Technical maturity is driven by cost-effectiveness. Despite differences in core goals, key technologies, scenarios, and technical status(Table 1), both domains require multimodal complementarity to overcome single-sensor limitations, reflecting cross-domain collaboration and enabling technology transfer (e.g., medical force control algorithms in industrial precision assembly).

Table 1: Horizontal comparison table

Comparison Dimensions	Medical Robots	Industrial Robots	Commonalities
Core Goals	Improving surgical precision and safety	Improving the efficiency and safety of human-machine collaboration	Balancing accuracy, real-time performance, and robustness
Key technologies	Tactile feedback (FBG/piezoelectric sensors), stereo vision, fluorescence imaging	Dynamic vision (event camera), lidar environment modeling, wearable devices	Multimodal complementarity needs
Typical scenarios	Robot-assisted suturing, minimally invasive biopsy	Dynamic obstacle avoidance, motor fault diagnosis	Rely on multi-modal collaboration to improve performance
Technology Status	Clinical validation stage (such as the Da Vinci system)	Large-scale deployment (e.g., collaborative robots)	Medical care focuses on ethical compliance, while industry focuses on cost and efficiency

DOI: 10.54254/2755-2721/154/2025.TJ23114

4.2. Future development direction

Advancements in medical technology necessitate deeper R&D into biocompatible sensors and dynamic compensation algorithms, like flexible electronics with self-healing structures, to enhance the stability of long-term implanted devices. Industrially, lightweight spatiotemporal fusion algorithms should be developed, strengthening event camera and lidar collaborative perception for optimized real-time decision-making.

Sensor redundancy optimization requires exploring dynamic modal selection mechanisms, hardware-algorithm co-design, and cross-domain redundancy governance strategies.

Cross-domain collaboration can leverage AI-driven adaptive fusion frameworks to overcome physical boundaries, enabling technical migration of force control algorithms and edge computing architectures.

Ethically and for standardization, dynamic privacy protection mechanisms and cross-industry security assessment systems are crucial to balance innovation and data compliance. Co-constructing new sensors and open-source ecosystems will promote technology adoption across sectors. Future research should emphasize multidisciplinary integration to advance adaptive algorithms and explore technology-ethics-industrialization collaboration, ensuring the reliability of intelligent robot systems.

5. Conclusion

Multimodal sensor fusion technology has significant application value in the fields of medical and industrial robots. In medical surgery, the combination of visual and tactile feedback improves the accuracy and safety of minimally invasive operations, and the introduction of a privacy protection framework ensures that medical data is legal and compliant. In industrial scenarios, event cameras and lidars work together to optimize the efficiency of dynamic obstacle avoidance, and proposal-level fusion methods significantly improve target detection performance through cross-modal feature interaction. However, the two major fields still face common challenges, such as breaking through the long-term stability of sensors and reducing the deployment cost of multimodal systems. It is worth noting that there is great potential for cross-domain technology interoperability. For example, stress compensation technology for medical tactile skin may be applied to industrial precision assembly scenarios, and industrial edge computing solutions are expected to accelerate the real-time processing efficiency of medical data. The combination of flexible tactile sensors in the medical field and industrial-grade dynamic obstacle avoidance algorithms may give rise to a new generation of adaptive robot systems.

References

[1] *Freschi, C., Ferrari, V., Melfi, F., Ferrari, M., Mosca, F., & Cuschieri, A. (2013). Technical review of the Da Vinci surgical telemanipulator. The International Journal of Medical Robotics and Computer Assisted Surgery, 9(4), 396-406.*

[2] *D'Ettorre, C., Mariani, A., Stilli, A., y Baena, F. R., Valdastri, P., Deguet, A., ... & Stoyanov, D. (2021). Accelerating surgical robotics research: A review of 10 years with the Da Vinci research kit. IEEE Robotics & Automation Magazine, 28(4), 56-78.*

[3] *McKinley, S., Garg, A., Sen, S., Gealy, D. V., McKinley, J. P., Jen, Y., ... & Goldberg, K. (2016, August). A n interchangeable surgical instrument system with application to supervised automation of multilateral tumor resection. In 2016 IEEE International Conference on Automation Science and Engineering (CASE) (pp. 821 -826). IEEE.*

[4] *Lee, Y. J., van den Berg, N. S., Orosco, R. K., Rosenthal, E. L., & Sorger, J. M. (2021). A narrative review of fluorescence imaging in robotic-assisted surgery. Laparoscopic surgery, 5, 31.*

[5] *Ryu, S. C., & Dupont, P. E. (2014, May). FBG-based shape sensing tubes for continuum robots. In 2014 IEEE International Conference on Robotics and Automation (ICRA) (pp. 3531-3537). IEEE.*

DOI: 10.54254/2755-2721/154/2025.TJ23114

[6] *Liu, Q., Dai, Y., Li, M., Yao, B., & Zhang, J. (2024). FBG-based sensorized surgical instrument for force measurement in minimally invasive robotic surgery. IEEE Sensors Journal.*

[7] *Naidu, A. S., Patel, R. V., & Naish, M. D. (2016). Low-cost disposable tactile sensors for palpation in minimally invasive surgery. IEEE/ASME Transactions On Mechatronics, 22(1), 127-137.*

[8] *Talasaz, A., Trejos, A. L., & Patel, R. V. (2016). The role of direct and visual force feedback in suturing using a 7-DOF dual-arm teleoperated system. IEEE transactions on haptics, 10(2), 276-287.*

[9] *Chen, S., Zhang, J., Cheng, N., Li, W., Liu, J., Xie, H., ... & Song, Y. (2025). Printing perovskite and graphene parallel structure-based optical-mechanical sensors for human-machine interaction. Science China Technological Sciences, 68(2), 1220201.*

[10] *Sanfilippo, F., & Pettersen, K. Y. (2015, November). A sensor fusion wearable health-monitoring system with haptic feedback. In 2015 11th International conference on innovations in information technology (IIT) (pp. 262-266). IEEE.*

[11] *Wang, J., Quasim, M. T., & Yi, B. (2025). Privacy-Preserving Heterogeneous Multi-Modal Sensor Data Fusion via Federated Learning for Smart Healthcare. Information Fusion, 103084.*

[12] *Mueggler, E., Rebecq, H., Gallego, G., Delbruck, T., & Scaramuzza, D. (2017). The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM. The International journal of robotics research, 36(2), 142-149.*

[13] *Kolar, P., Benavidez, P., & Jamshidi, M. (2020). Survey of datafusion techniques for laser and vision based sensor integration for autonomous navigation. Sensors, 20(8), 2180.*

[14] *Neto, P., Simão, M., Mendes, N., & Safeea, M. (2019). Gesture-based human-robot interaction for human assistance in manufacturing. The International Journal of Advanced Manufacturing Technology, 101, 119-135.*

[15] *Chen, M., Liu, C., & Du, G. (2018). A human–robot interface for mobile manipulator. Intelligent Service Robotics, 11, 269-278.*

[16] *Castanedo, F. (2013). A review of data fusion techniques. The scientific world journal, 2013(1), 704504.*

[17] *Chandan, K., Zhang, X., Albertson, J., Zhang, X., Liu, Y., & Zhang, S. (2020, July). Guided 360-degree visual perception for mobile telepresence robots. In The RSS-2020 Workshop on Closing the Academia to Real-World Gap in Service.*

[18] *Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).*

[19] *Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence, 27(8), 1226-1238.*

[20] *Hamaide, V., & Glineur, F. (2021). Unsupervised minimum redundancy maximum relevance feature selection for predictive maintenance: Application to a rotating machine. International Journal of Prognostics and Health Management, 12(2).*

[21] *Zugasti, E., Mujica, L. E., Anduaga, J., & Martinez, F. (2013). Feature selection-Extraction methods based on PCA and mutual information to improve damage detection problem in offshore wind turbines. Key Engineering Materials, 569, 620-627.*

[22] *Jing, L., Wang, T., Zhao, M., & Wang, P. (2017). An adaptive multi-sensor data fusion method based on deep convolutional neural networks for fault diagnosis of planetary gearbox. Sensors, 17(2), 414.*

[23] *Kim, Y., Kim, S., Choi, J. W., & Kum, D. (2023, June). Craft: Camera-radar 3d object detection with spatio-contextual fusion transformer. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 1160-1168).*