Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 8, No. 2, 2026

Reflections on Practical Paths of Empowering Embodied AI Development via Computer Vision in Robotic Systems

Zichen Wang

School of Sussex University, Brighton, UK

Abstract: Currently, most commercial robots rely on fixed programs to perform repetitive tasks. The visual modules equipped on these devices have limited anti-interference capabilities, making it difficult to adapt to complex structures and variable environments, which in turn limits the practical effectiveness of robotic intelligence deployment. Drawing from hands-on experience in robot debugging and project implementation, this paper explores the integration of computer vision and embodied intelligence. Based on common application scenarios such as industrial sorting, power inspection, and intelligent services, it analyzes specific methods by which visual technologies assist robots in environmental perception, intelligent decision-making, motion adjustment, and model updating. The study identifies several common challenges encountered during industry deployment, including insufficient stability in recognizing complex scenes, difficulty balancing computational power with recognition accuracy, significant gaps between simulation training and real-world conditions, and a lack of unified application standards across industries. Addressing these practical issues, the paper proposes feasible improvement strategies from four perspectives: algorithm enhancement, computational power upgrade, simulation optimization, and the establishment of industry standards. The conclusions drawn from this research offer practical insights for advancing robot vision technology, expanding application scenarios, and promoting standardized development within the industry.

Keywords: Robotic system, computer vision, Embodied Artificial Intelligence, Perceptual Interaction, Intelligent empowerment.

1. Introduction

At present, the working modes of most industrial and service robots are relatively fixed. The equipment follows preset trajectories and fixed instructions to complete standardized operations. The advantages of this operation mode are very prominent, with stable operation status and high controllability. It can adapt to simple scenarios with regular factory and warehouse environments. This is also the main reason why industrial robots have been able to be widely adopted quickly. However, in real engineering application scenarios, the shortcomings of this fixed operation mode will be fully exposed. There are a large number of uncertain factors in the actual operation site, such as changes in light intensity, random material obstruction, and temporary changes in working conditions. Traditional robots cannot autonomously respond to these sudden changes and often experience positioning deviations, grasping errors, and emergency shutdowns. Due to these problems, traditional robots can only complete basic repetitive tasks and have difficulty achieving intelligent upgrades, and are difficult to adapt to the increasingly complex operational requirements at present [1].

Traditional automated equipment only executes preset instructions passively. The core advantage of embodied intelligent robots lies in their ability to actively interact with the physical environment and adjust their operation status according to the actual situation on site, completely changing the rigid operation logic of robots. Embodied intelligent robots can accumulate a large amount of scene data during continuous operation, autonomously adjust equipment operating parameters, gradually break free from the constraints of fixed programs, and ultimately achieve autonomous adaptation to scenarios and independent completion of operational tasks. In the entire embodied intelligent working system, computer vision is the core tool for robots to perceive the external environment, responsible for basic tasks such as image acquisition, feature extraction, and scene reconstruction, effectively compensating for the single perception form and stiff human-machine interaction of traditional robots [2]. Even though visual technology has obvious advantages in theory, when combined with the actual feedback from a large number of implemented projects, various environmental disturbances in complex sites will significantly reduce the stability of the algorithm, and the theoretical technical advantages are difficult to be fully implemented in actual operations.

The rapid development of the intelligent manufacturing and intelligent service industries has broadened the application boundaries of robots. Equipment is no longer limited to standardized industrial factories and is gradually used in complex scenarios such as field inspections and handling of irregular materials. The operational requirements for robots have also continued to improve, not only focusing on basic operational efficiency, but also emphasizing the precision operation of equipment, autonomous adaptation, and unmanned work capabilities. Relying on constantly updated visual technology, robots have broken through the limitations of traditional static recognition and have acquired intelligent capabilities such as scene semantic judgment, motion trajectory prediction, and environmental pattern analysis. In the first-line debugging work, it can be clearly observed that high-quality algorithm models trained in the laboratory are easily affected by external factors such as light, obstruction, and dust when put into real complex scenarios, resulting in recognition deviations, decision delays, and other problems [3]. The incompatibility between the perception and decision-making modules, and the low adaptability between simulation training and actual deployment, have also greatly restricted the wide-scale promotion of visual embodied intelligent technology. This article, based on the practical experience accumulated from project operations, sorts out the actual logic of visual technology empowering embodied intelligent, summarizes the existing industry pain points, and proposes corresponding optimization ideas.

2. The Coupling Logic Between Computer Vision and Embodied Intelligence of Robots

Compared with traditional automated equipment, the most distinctive feature of embodied intelligent robots is their ability to adapt to scenarios and optimize data on their own. These robots rely on five modules - perception, cognition, decision-making, execution, and iteration - to form a complete working loop [4]. They continuously optimize operational parameters in environmental interactions and achieve an upgrade from mechanical repetitive tasks to autonomous intelligent operations. Currently, the common environmental perception methods for robots include laser perception, ultrasonic perception, and visual perception. Each technology has its own advantages and applicable scenarios. There is currently no universal perception solution that can adapt to all working conditions in the industry.

After years of development, laser perception and ultrasonic perception technologies have matured and have stable measurement ranges, suitable for short-distance distance measurement and basic obstacle avoidance in simple operation scenarios. However, the functions of these two technologies are relatively limited, only capable of completing basic distance detection and cannot recognize scene content, target shape, spatial layout, etc., which are deep-level information, and are insufficient to support high-precision intelligent decision-making requirements in complex scenarios [5]. Computer vision technology has more obvious comprehensive advantages, with lower hardware deployment costs and a wider range of applicable scenarios. The visual technology can simultaneously collect multi-dimensional data such as images, spatial structure, and scene semantics, providing sufficient data support for the robot to understand the environment and make autonomous decisions. Many on-the-ground cases of robot intelligent transformation prove that introducing computer vision technology is a key way to break through the traditional fixed operation mode and improve the intelligence level of equipment.

The accuracy and operational stability of robots are largely determined by the operating state of the visual perception system. The visual system is equivalent to the robot's visual sense, directly affecting the on-site operation quality of the equipment. With the support of visual technology, robots can complete fine operations such as target segmentation, posture judgment, and three-dimensional scene modeling, accurately capture real-time changes in the on-site environment, and achieve intelligent functions such as autonomous obstacle avoidance, work condition adaptation, and human-machine collaboration, eliminating the mechanical and passive operation shortcomings of traditional robots [6]. With the continuous upgrading of industry production demands, new visual technologies such as multi-dimensional feature fusion and dynamic scene modeling have gradually been put into application, effectively improving problems such as slow response, weak scene adaptability, and low operational efficiency of traditional robots, making the operation performance of intelligent robots more in line with complex and variable actual working conditions.

3. The Core Practical Approach of Endowing Robots with Embodied Intelligence Through Computer Vision

The empowerment effect of visual technology on embodied intelligent robots covers the entire operation process, running through core links such as environmental cognition, intelligent decision-making, action execution, and model update. Various functional modules of the robots cooperate and operate in coordination, building a complete intelligent operation system, enabling the equipment to handle various complex and dynamic engineering operation scenarios with ease.

Stable and precise environmental perception is the foundation for robots to achieve unmanned autonomous operations. Using either two-dimensional vision or three-dimensional sensing technology alone will have obvious shortcomings and is difficult to simultaneously balance recognition accuracy and spatial positioning effect [7]. A large number of on-site debugging data show that the combined perception scheme of integrating two-dimensional vision with three-dimensional point clouds has the best comprehensive performance. The two types of technologies can complement each other's advantages, and the equipment can flexibly adjust its working mode according to the complexity of the scene and reasonably allocate the computing power of the equipment. In standardized factory buildings and warehouses, robots can quickly complete scene modeling and target positioning; in dynamic and complex scenarios such as field inspection and messy material sorting, they can also real-time lock on the operation target, predict the movement trend of objects, and respond to various unexpected disturbances. This technical solution has been widely applied in industrial quality inspection, power inspection, intelligent sorting and other fields, effectively enhancing the robot's scene adaptability and operational stability [8].

Early intelligent robots often had problems of module fragmentation, with the perception unit and the decision-making unit operating independently, and there was a delay in data transmission and information synchronization, which was an important reason for frequent decision-making errors and running stalling of the equipment. Cross-modal fusion technology has well solved this industry problem. The visual language model can integrate multi-source information such as on-site images, manual operation instructions, and equipment operation status to autonomously complete semantic parsing and operation analysis. During human-machine collaborative operations, the robot can use visual recognition to identify the operator's operation posture and intention, actively adapt to the human work rhythm, and achieve safe and efficient collaborative cooperation. Combined with three-dimensional sensing and physical modeling technology, the robot can accurately judge the target operating status, identify potential safety hazards on site, and steadily improve the decision-making accuracy rate, breaking the complete operation loop from environmental perception to equipment execution [9].

The accuracy of operation and the stability of operation are the core indicators for evaluating the practical application value of robots. The operation accuracy of traditional robots is entirely dependent on mechanical hardware structure, with very little tolerance space. Even a slight environmental fluctuation or target position deviation can cause operation failure and cannot meet the requirements of high-precision operations such as precise assembly and grasping of irregular materials. The visual servo control system effectively compensates for this hardware defect. The system uses real-time visual feedback to complete dynamic adjustment and continuously corrects the errors caused by mechanical operations, making up for the inherent deficiencies of the hardware structure. In precise assembly operations, this technology can achieve sub-millimeter-level high-precision calibration, even when encountering target displacement, temporary obstruction, and other unexpected situations, it can quickly adjust the operation trajectory, through the efficient cooperation of perception, decision-making, and execution, to meet the strict precision standards of precise operations [10].

The general operation ability and multi-scenario adaptability of robots need to be continuously optimized and upgraded through massive and diverse scenario data. In real engineering scenarios, the sample collection of special working conditions such as extreme lighting, complex obstructions, and target deformation is difficult and costly. Training only with real scene data is difficult to fully optimize the comprehensive performance of the model. The virtual-real iterative training technology solves this industry pain point. The simulation system can generate various complex and extreme working condition data, enriching the training dimensions of the model, and significantly reducing the human and time costs of real scene debugging. The industry generally adopts the virtual-real iterative optimization mode, continuously improving the model's generalization ability and reducing on-site failure problems of the robot. After years of industry technology iteration, this virtual-real combination training system has become mature, providing strong support for the large-scale implementation of embodied intelligent technology [11].

4. The Current Technical Bottlenecks and Implementation Issues in the Empowerment Process

The enabling value of visual technology for embodied intelligence has been widely recognized in the industry. Various related application solutions have gradually shifted from the laboratory research stage to the industrialization implementation stage. However, the actual situations of on-site debugging and project operation reveal that the technical advantages under the ideal laboratory environment are difficult to directly translate into complex real operational scenarios. Multiple problems such as the technical shortcomings, the difficulty in scene adaptation, and the incomplete industry supporting system have cumulatively slowed down the overall process of technology's large-scale implementation. The core issues currently existing in the industry mainly lie in four aspects [12].

The instability of perception in complex scenarios is the most prominent practical problem in the industry's implementation process. The training environments of mainstream visual models are all laboratory pure datasets, with a single scene and very little interference, which is greatly different from the real and complex engineering sites. In common operation scenarios such as outdoor power inspection and messy industrial workshops, sudden changes in light, rain and snow cover, stacked materials, and missing target textures frequently occur. Various external interferences directly reduce the model's anti-interference ability, causing feature failure, target omission, and recognition misalignment. Such perception errors will directly affect the operation accuracy and stability of robots, and are the core factor restricting the large-scale commercial application of visual intelligent technology [13].

The conflict between recognition accuracy and equipment response speed has long plagued the on-site implementation of visual algorithms. High-precision cross-modal algorithm operation logic is cumbersome and consumes a large amount of computing power. Most lightweight terminal devices cannot support the real-time stable operation of the algorithm. To adapt to the computing power and power consumption limitations of terminal devices, the industry mostly simplifies the algorithm structure to reduce computing power loss, but this optimization method will lose a large number of detailed features, directly causing a decrease in recognition accuracy. During actual debugging, it can be clearly observed that the response speed of the equipment and the recognition accuracy cannot be balanced, and the balance relationship between the two has always been difficult to break. The problem of insufficient soft and hardware compatibility between perception and control modules will also trigger operational errors, sudden equipment stops, and other chain failures, seriously affecting the stability of continuous robot operations.

The problem of deviation in the adaptation of virtual and real scenarios cannot be completely eliminated, and has long affected the real-world implementation effect of the model. The simulation platform is the core tool for model development and parameter debugging, which can effectively shorten the development cycle and reduce the debugging cost. However, the scene reproduction ability of existing simulation systems is limited, and cannot fully replicate the texture, lighting, physical characteristics, and various random interferences of the real scene. Many models that perform well in simulation tests will have a significant performance decline when deployed on physical robots, resulting in scenarios adaptation failure, decision-making deviations, and accuracy attenuation and other various problems. The existing optimization methods can only slightly narrow the gap between virtual and real scenarios, and cannot fundamentally solve the adaptation problem. This not only prolongs the project debugging cycle but also increases the overall cost of technology implementation [14].

In addition to various technical bottlenecks, the incomplete industrialization support and poor technical universality of the industry also limit the speed of the popularization of visual embodied intelligent technology. Currently, the industry mostly adopts a single-scenario customized research mode, developing exclusive models for specific scenarios, with weak cross-scenario adaptation ability and low reusability. This customized development mode not only has high investment costs and long research and development cycles, but also makes it difficult for the industry to form unified technical standards and implementation norms. High-precision sensors, high-end computing power chips, and other core hardware are priced relatively high, raising the threshold for intelligent upgrades of small and medium-sized enterprises. There are very few mature and universal solutions for adapting to ordinary low-cost equipment, leaving high-quality visual intelligent technology limited to high-end industrial scenarios and unable to achieve full industry-wide popular application.

5. Optimization Strategies for the Development of Embodied Intelligence Enabled by Computer Vision

Based on the technical shortcomings and industry pain points exposed in the practical engineering operations, and in line with the overall development trend of the intelligent robot industry, this article systematically outlines a set of optimized solutions from four dimensions: algorithms, computing power, simulation, and industrial implementation. These solutions are targeted at solving various implementation problems and help visual embodied intelligent technology develop towards large-scale and inclusive directions [15].

The core idea of algorithm optimization is to break through the training limitations of pure laboratory data sets and build a diversified training sample library based on real production scenarios. Staff can collect real scene data including light interference, material occlusion, and extreme conditions by combining various engineering sites, and combine data enhancement, dynamic feature selection, etc. to improve the model's anti-interference ability and scene adaptability. The promotion of the dual-mode perception architecture combining two-dimensional vision and three-dimensional point cloud in the industry, combined with the adjustment of model reasoning logic according to the physical operation laws of the real scene, can effectively improve the recognition accuracy and operational stability in complex conditions, and alleviate the common problem of unstable perception in complex scenarios.

The optimization and upgrade of the computing power architecture can effectively balance the recognition accuracy and device response speed, and solve the common problem of insufficient computing power of terminal devices. Fine-tuning the cross-modal basic model, eliminating redundant parameters and simplifying reasoning logic, can reduce computing power consumption while retaining the core recognition performance, and adapt to the operating conditions of lightweight terminal devices. The hierarchical computing power system of end-cloud collaboration can reasonably allocate device tasks, with the terminal responsible for real-time perception and dynamic control, and the cloud undertaking complex reasoning, model updates, and big data analysis, and by optimizing data transmission links, balance the operation accuracy of the equipment and the overall operational efficiency [16].

The optimization of the simulation system is mainly to narrow the migration gap between virtual and real scenarios and improve the success rate of model deployment in real scenes. Researchers can upgrade the physical modeling capability of the simulation platform to accurately restore the core elements such as lighting, texture, and mechanical properties of the real scene. By setting random differentiated parameters, the types of simulation conditions can be enriched, and the training coverage of the model can be expanded. The introduction of frontier technologies such as small sample learning and adaptive domain adaptation can enhance the model's cross-scenario generalization ability, reduce the manpower and time investment in real scene debugging, and make the operation effect of the model after implementation more stable.

In the industrial implementation stage, the inefficient single-scenario customized R&D mode needs to be abandoned. The industry R&D team can focus on researching multi-condition universal basic models to reduce overall R&D costs and promote the industry to form unified technical implementation standards. Continuous optimization of the compatibility between lightweight algorithms and low-cost visual hardware can effectively improve the overall solution's cost performance [17]. By combining the scene characteristics of different fields such as industrial production, intelligent services, and field inspection, corresponding standardized implementation rules can be formulated to lower the threshold for intelligent upgrading of small and medium-sized enterprises and promote the large-scale, low-cost popularization and application of visual embodied intelligent technology.

6. Conclusion

Based on the practical experience of on-site robot debugging and project implementation, this article systematically explores the internal logic of how visual technology enables embodied intelligent robots, and builds an integrated intelligent empowerment system covering perception, reasoning, execution, and iteration. Different from pure theoretical research, this article combines real industrial scenarios such as industrial sorting and power inspection, and clarifies the practical value of visual technology in improving the drawbacks of traditional robot programmatic and rigidized operations. The article summarizes various implementation challenges such as insufficient perception stability, imbalance in computing power and accuracy, large deviations in virtual-real adaptation, and the absence of industry standards. It also provides comprehensive optimization solutions, which can provide solid practical references for the iteration of robot vision technology and the large-scale application of scenarios.

This research has certain limitations. Overall, it mainly relies on experience summary and qualitative analysis, without conducting multi-condition control experiments, and lacks quantitative comparison data on the performance of different visual algorithms. It is difficult to accurately determine the actual improvement effect of various optimization schemes. At the same time, the article's discussion on the customized implementation details of each industry is insufficient, and the proposed optimization schemes are relatively general and have limited scene specificity. There is still considerable room for further refinement and improvement.

In response to the shortcomings of this research, future work can focus on supplementing quantitative test content, conducting multi-condition performance tests to accurately verify the actual application effects of various optimization strategies. Focusing on core industry pain points such as virtual-real scene adaptation deviation and lightweight general model research and development, continuously optimize the simulation training system to narrow the performance gap between theoretical training and real-world deployment. Continuously improve the technical system and industry implementation norms, enhance the environmental adaptability and autonomous operation level of robots, and promote the standardized and large-scale implementation of visual embodied intelligent technology, helping the domestic intelligent robot industry achieve high-quality development.

References

[1] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., et al. (2017). Building machines that learn and think like people. Behavioral and Brain Sciences, 40, e253.

[2] Howard, D., Eiben, A. E., Kennedy, D. F., et al. (2019). Evolving embodied intelligence from materials to machines. Nature Machine Intelligence, 1(1), 12–19.

[3] Jin, D., & Zhang, L. (2020). Embodied intelligence weaves a better future. Nature Machine Intelligence, 2(11), 663–664.

[4] Du, S., & Fei, Y. (2022). Embodied artificial intelligence: From perception to physical interaction. Nature Machine Intelligence, 4(11), 927–934.

[5] Cadena, C., Carlone, L., Carrillo, H., et al. (2016). Past, present, and future of simultaneous localization and mapping: toward the robust-perception age. IEEE Transactions on Robotics, 32(6), 1309–1332.

[6] Lepetit, V., & Fua, P. (2015). Monocular 3D tracking of rigid and articulated objects. International Journal of Computer Vision, 111(2), 147–170.

[7] Bohg, A., Morales, A., Asfour, T., et al. (2014). Data-driven grasp synthesis—a survey. IEEE Transactions on Robotics, 30(2), 289–309.

[8] Yang, S., & Scherer, S. (2019). Direct monocular odometry with active feature selection for robust navigation. IEEE Transactions on Robotics, 35(4), 963–979.

[9] Zhang, J., & Singh, S. (2018). LOAM: lidar odometry and mapping in real-time. IEEE Transactions on Robotics, 34(4), 1056–1070.

[10] Ren, L., Dong, J., Liu, S., et al. (2025). Embodied intelligence toward future smart manufacturing in the era of AI foundation model. IEEE/ASME Transactions on Mechatronics, 30(4), 2632–2642.

[11] Hussein, A., Gaber, M. M., Elyan, E., et al. (2017). Imitation learning: a survey of learning methods. ACM Computing Surveys, 50(2), 1–35.

[12] Zhang, D. (2022). Explainable hierarchical imitation learning for robotic drink pouring. IEEE Transactions on Automation Science and Engineering, 19(4), 3871–3887.

[13] Levine, S., Finn, C., Darrell, T., et al. (2016). End-to-end training of deep visuomotor policies. Journal of Machine Learning Research, 17, 1–40.

[14] Pinto, L., & Gupta, A. (2017). Supersizing self-supervision: learning to grasp from 50k tries and 700 robot hours. The International Journal of Robotics Research, 36(10-11), 1231–1250.

[15] Zhu, Y. (2020). Target-driven visual navigation in indoor scenes via deep reinforcement learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(10), 2409–2423.

[16] Shridhar, M., Manuelli, L., & Fox, D. (2023). CLIPort: what and where pathways for robotic manipulation. The International Journal of Robotics Research, 42(1-2), 112–130.

[17] Liu, X., Li, Y., & Wang, H. (2026). Vision-language-action foundation models for robotic embodied intelligence: survey and open challenges. IEEE Transactions on Cybernetics, 56(2), 789–804.