Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/150/2025.22708

The Evolution and Convergence of Artificial Intelligence and Intelligent Robotics

Minghao Hong

Beijing University of Posts and Telecommunications, Beijing, China
harsuz@sina.com

Abstract: Nowadays, AI's development, especially the emergence of large models, has given it a depth of thinking and generalization ability it never had before, and this ability can be applied across the board. As a field that has long intersected with AI technology, intelligent robots have shifted from past modes of thought and technical routes to today's large models. This paper comprehensively traces the technical evolution of intelligent robots from their inception to now, focusing on their path of integration with AI. Using literature research, comparative analysis, and trend analysis, combined with in - depth technical analysis, it identifies key moments in the evolution of intelligent robot technology and the internal logic of its integration with AI. The research indicates that the convergence of artificial intelligence and robotics has progressed from mere perception to cognitive capabilities, transitioning from specialized to general-purpose applications. Moving forward, this integration is expected to become increasingly prevalent as technological advancements steer towards broader applications.

Keywords: AI Robot, Multimodality, Artificial Intelligence, Reinforcement Learning, large language model

1. Introduction

The integration of artificial intelligence (AI) and robotics have undergone a transformative evolution, propelled by advancements in machine learning paradigms. Historically, robots relied on model-based controllers and hand-coded rules, which proved inadequate for dynamic environments and high-dimensional perceptual-motor tasks. Early breakthroughs in reinforcement learning (RL) demonstrated its potential for adaptive decision-making, yet limitations remained in training efficiency and generalization to real-world scenarios. Subsequently, the development of deep reinforcement learning (DRL) and multimodal perception filled these gaps, enabling robots to process visual, tactile, and language inputs in a coordinated manner.

The advent of large language models (LLMs and VLMs) has transformed the trajectory of AI and robotics, enhancing cross-modal reasoning and task generalization. This study examines the integration of AI and robotic technologies, highlighting the transition from reinforcement learning to multimodal frameworks and large-scale model-driven embodied intelligence, which redefines robot autonomy. By analyzing the developmental continuum, this review systematically addresses the relationship between AI and robotics, focusing on the convergence of ideas and the rise of general embodied intelligence in the context of large models. Key components of this evolution include scalable parallel reinforcement training, cross-modal knowledge transfer, integrated

DOI: 10.54254/2755-2721/150/2025.22708

representation learning, and end-to-end models. The integration of these technologies has led to a new paradigm of large language models, enhancing the adaptability of general-purpose robots in open environments. This underscores the need for standardized benchmark testing to assess emerging capabilities like zero-shot task transfer. Theoretically, a reexamination of embodiment in AI systems is warranted, viewing large-scale models as a universal prior for physical interaction. The advanced comprehension and generalization abilities of large models should be leveraged in robotics to further investigate general intelligence.

2. Early stage-application of reinforcement learning in robotic decision-making

2.1. Reinforcement learning and robotic decision-making

Initially, the intersection of robot control and artificial intelligence were limited, with controllers relying heavily on precise models and manual design. However, as the complexity of task scenarios and interactions increased, the limitations of traditional control methods gradually became apparent [1]. Model-based control struggles to cope with real-world complex scenarios, and the high dimensionality of sensor data and continuous actions can lead to the curse of dimensionality. Traditional tasks such as navigation and grasping often have sparse reward signals, necessitating long-term planning. Reinforcement learning, on the other hand, can optimize control strategies through a trial-and-error mechanism and possesses adaptability, reactivity, and self-supervision characteristics. Compared with traditional manually designed models, reinforcement learning architectures leverage neural networks to significantly enhance learning speed and generalization ability [1].

Reinforcement learning (RL) provides a method for learning optimal policies through trial and error. By leveraging neural networks as powerful function approximators, it can handle high-dimensional data inputs and learn complex nonlinear mappings, thus addressing the challenges faced by traditional RL in high-dimensional state and action spaces [1]. This breakthrough deviates from the conventional approach of manually designing controllers. It demonstrates the potential of combining neural networks with reinforcement learning in robotic control, offering new insights for adaptive learning in robots.

2.2. Further development of reinforcement learning-based decision-making

With the idea of integrating reinforcement learning frameworks into robot decision-making control, the convergence of artificial intelligence and robotics technologies has become increasingly tight.

Deep reinforcement learning has achieved remarkable success in many robot decision-making tasks [1], but the training process typically requires substantial time and computational resources. Moreover, traditional deep reinforcement learning methods (such as DQN and A3C) rely on synchronous updates, which leads to low overall training efficiency. The subsequent emergence of asynchronous reinforcement learning frameworks, combined with distributed computing and parallelization techniques, has significantly accelerated the training process and enhanced algorithm performance [2]. Asynchronous reinforcement learning showcases the potential for parallelization in deep reinforcement learning, establishing a basis for large-scale distributed approaches. The A3C algorithm has emerged as a seminal method, catalyzing extensive subsequent research.

Although improvements have been made in algorithm speed and resource efficiency, for robotic manipulation tasks such as grasping, placing, and assembly, high-precision control and complex decision-making capabilities are usually required. Deep reinforcement learning has demonstrated the potential of artificial intelligence in robotic manipulation control tasks [2]. It has proposed a deep reinforcement learning framework suitable for high-dimensional sensor data and continuous action spaces, essentially addressing the limitations of traditional methods.

DOI: 10.54254/2755-2721/150/2025.22708

3. Mid-stage-from deep reinforcement learning to multimodality

3.1. The rapid development and new limitations of deep reinforcement learning

Deep learning has also achieved unprecedented results [3] in core decision-making challenges such as autonomous navigation. Autonomous navigation is a fundamental challenge in robotics, particularly in complex and dynamic environments. Conventional methods depend on accurate maps and predefined rules, but they struggle with dynamic obstacles and uncertainties [4]. Deep reinforcement learning (Deep RL) provides a method for learning optimal strategies through trial and error, enabling the direct learning of navigation strategies from sensor data.

Mirowski, P. proposed a navigation framework based on deep reinforcement learning, which directly learns navigation policies from raw sensor data, such as LiDAR and cameras. By integrating Deep Q-Networks (DQN) and the Actor-Critic architecture, he designed a reinforcement learning algorithm suitable for complex environments [4]. The remarkable performance of this framework demonstrates the potential of deep reinforcement learning in navigation within complex environments, offering new insights for autonomous navigation.

It can be observed that the adaptation and integration of reinforcement learning at the decision-making layer of robots have revealed the potential of artificial intelligence in decision-making. However, artificial intelligence can be applied not only at the decision-making layer of robots. The integration of artificial intelligence into the original sensor perception layer of robots can also bring performance improvements that are different from those of the past, and can complement the decision-making layer based on reinforcement learning. This paves the way for the transition from deep reinforcement learning technology to multimodal technology.

Traditional computer vision methods depend on extensive labeled datasets for model training, which presents significant challenges, including high costs and time demands. In contrast, humans acquire visual information through observation and interaction without explicit labeling. This suggests the potential for a self-supervised learning framework utilizing inter-frame motion data in video sequences, such as camera motion, as a supervisory signal. By predicting camera motion (e.g., rotation and translation), the model can learn meaningful visual representations [5]. This approach reduces the dependence on external labeled data and demonstrates the potential of motion signals in visual representation learning.

3.2. Simulation training and real-world interaction

Before the development of multimodal thinking, researchers had already observed that as the requirements for task design and machine data volume increased, the training of robot models became a significant issue. The emergence of the Sim-to-Real concept addressed this problem that spans model training and interaction with the physical world.

Traditional methods of decision-making and control relied on precise dynamic modeling or domain adaptation, but these approaches faced limitations in complex real-world environments. The Sim-to-Real concept aims to facilitate robot control transfer from simulation to reality through dynamic randomization. This allows for simulation training of specific real-world scenarios. However, due to inherent dynamic discrepancies, strategies developed in simulation often do not transfer effectively to real robots. However, the outcome is entirely different if dynamic randomization is introduced by incorporating randomized dynamic parameters in the simulation [6]. Such an operation not only enhances the robustness of the strategies but also enables them to adapt to the dynamic changes in the real world.

The augmentation of dynamics randomization has demonstrated the potential for policies trained in simulation to be directly applied to real-world robots [6], thereby reducing the cost and risk

DOI: 10.54254/2755-2721/150/2025.22708

associated with physical robot experiments. This advancement has propelled the development of subsequent robot models that require larger datasets, face more complex scenarios, and encounter increased training difficulties.

3.3. The concept of multimodal fusion

In robotic manipulation tasks like grasping, placing, and assembling, precise object perception and localization are essential. While deep reinforcement learning enables autonomous decision-making, a robust perception system is crucial for effective manipulation. Traditional perception methods, which depend on hand-crafted features or geometry-based models, often struggle with complex shapes, textures, and deformable objects. Dense Object Nets (DON) is a method for robotic manipulation tasks that achieves this by learning dense visual object descriptors [7]. Dense visual descriptors provide semantic information for every pixel on the object's surface, offering a richer perception capability for robotic manipulation tasks. The integration of this enriched perception capability with deep reinforcement learning demonstrates the potential of combining perceptual data with self-supervised learning [7]. This, in turn, has inspired subsequent multimodal approaches.

The concept of multimodal perception originated in grasping tasks. Robotic grasping tasks require precise sensing and control capabilities, especially when dealing with complex shapes, fragile objects, or dynamic environments. Conventional approaches predominantly depend on visual data; however, upon making contact with an object, visual feedback alone may prove inadequate for conveying essential information. In contrast, tactile data can yield critical insights, including contact force, object slippage, and surface properties, thereby introducing an additional perceptual dimension crucial for grasping and regripping activities.

A multimodal perception framework that integrates vision and touch is employed for robotic grasping and regripping tasks [8]. This framework utilizes deep neural networks to process visual information (such as images) and incorporates tactile sensor data (such as force and pressure distribution) as supplementary inputs. A multimodal fusion strategy has been designed to combine visual and tactile information, thereby enhancing the success rate and stability of grasping [8].

The excellent performance achieved by the integration of multimodalities [8] demonstrates the potential of combining vision and touch in robotic grasping tasks, offering new insights for multimodal perception. Moreover, a multimodal fusion framework suitable for high-dimensional sensor data and complex tasks has been proposed, addressing the limitations of traditional methods.

4. Modern stage-empowerment by large language models and embodied intelligence

4.1. The general capabilities of large language models

Mirchandani et al. pointed out that large language models (LLMs) possess the ability for cross-modal pattern recognition [9]. The cross-modal interaction, control, and decision-making capabilities of large language models, like GPT-3 and BERT, enhance their compatibility with robotic decision-making. Their robust reasoning and generalization abilities enable powerful pattern recognition and generation across various domains, including natural language processing, image processing, time-series analysis, and symbolic reasoning. Exploring large language models as universal pattern machines may yield integrated solutions for cross-domain applications [9].

For robots, large language models can serve as a universal pattern, offering a unified solution for the acquisition of multimodal information, decision-making, and control processes required by robots.

DOI: 10.54254/2755-2721/150/2025.22708

4.2. Large language models and general-purpose intelligent robots

Brohan et al.'s RT-1 represents a pioneering application of the Transformer architecture in real-world robotic control, adeptly processing multimodal inputs, including images and states, to generate low-level control commands. This model is engineered to facilitate efficient learning and execution of extensive real-world robotic tasks. In addition, RT-2 enhances this framework by incorporating visual-language pre-trained models, such as CLIP, to facilitate the transfer of internet-derived knowledge into robotic control applications. Both RT-1 and RT-2 operate within the Visual-Language-Action (VLA) model framework for robotic control. RT-2 advances the capabilities of RT-1 by utilizing large-scale visual-language pre-trained models, including CLIP and Florence, as a foundational element, thereby enabling the fine-tuning process to effectively transfer internet knowledge to robotic control tasks [11]. The model is capable of directly generating robot control commands from visual inputs and language instructions, achieving end-to-end learning and execution [11].

It is not only the RT-2 based on the Transformer architecture that can be utilized, but also the text-generation capability of large language models (LLMs) themselves as a direct channel for generating control instructions. Liang et al. proposed using LLMs to generate executable code to control robots [12]. Code as Policies (CaP), a framework that employs code generated by language models as control policies, is used to implement embodied control tasks. By leveraging the code generated by language models as control policies, it achieves an end-to-end mapping from natural language instructions to robot actions [12].

Ahn et al. introduced the concept of Affordances. Robot control tasks typically require the translation of natural language instructions into concrete actions. However, language instructions often fail to fully capture the actual capabilities of robots (Affordances) [13]. Traditional methods rely on precise language understanding and action mapping, but they perform poorly when facing complex tasks and dynamic environments. By integrating language instructions with the functional capabilities of robots, more natural and efficient control can be achieved [13]. This is enabled by leveraging the general capabilities of large models, combined with multimodal perception, to create end-to-end intelligent decision-making robots.

4.3. Embodied intelligence

Original embodied intelligence has gained developmental insights through large models. Gupta et al. proposed optimizing embodied agents by integrating reinforcement learning with evolutionary algorithms. This approach highlights that agents learn and evolve via environmental interaction, rather than relying on fixed rules. Traditional methods often apply learning (e.g., reinforcement learning) or evolution (e.g., genetic algorithms) in isolation, which is inadequate for complex, dynamic environments. Combining these strategies harnesses their strengths, leading to more efficient adaptive behaviors [14].

PaLM-E, an embodied multimodal language model, aims to integrate vision, language, and robot control into a unified framework to create more intelligent embodied agents. It combines multimodal capabilities, large language models, and embodied intelligence [15]. PaLM-E integrates vision, language, and robot control into a unified framework, using a pre-trained large language model (such as PaLM) as the foundation and extending its capabilities to handle multimodal inputs (such as images and sensor data). A multimodal fusion mechanism is designed to combine visual and linguistic information to generate robot control commands [15].

DOI: 10.54254/2755-2721/150/2025.22708

5. Conclusion

The integration of AI and robotics has progressed from isolated perceptual modules to cohesive embodied intelligent systems. Early reinforcement learning frameworks established a basis for adaptive decision-making but encountered computational limitations. The advent of deep reinforcement learning and asynchronous training architectures, such as A3C, mitigated these challenges, allowing robots to manage high-dimensional states in manipulation and navigation tasks. Nonetheless, dependence on curated datasets and rigid simulation environments constrains real-world applicability. The simulation-to-reality paradigm, bolstered by dynamic randomization, has emerged as a crucial solution, enhancing policy robustness across various domains.

Multimodal learning represents a significant advancement. Frameworks like Dense Object Nets (DON) and visual-tactile fusion illustrate that cross-sensory integration markedly improves task success rates, paving the way for large-scale models. The rise of large language models subsequently redefines general-purpose robots through three main strategies: cross-modal knowledge transfer (e.g., RT-2 utilizing Vision-Language Models to align internet-scale data with control policies), code generation as executable plans (e.g., "code as policy"), and a unified representation space (e.g., PaLM-E's joint embedding of language, vision, and kinematics).

References

[1] *Lin, L. J. (1992). Reinforcement learning for robots using neural networks. Carnegie Mellon University.*

[2] *Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... & Kavukcuoglu, K. (2016, June). Asynchronous methods for deep reinforcement learning. In International conference on machine learning (pp. 1928-1937). PmLR.*

[3] *Gu, S., Holly, E., Lillicrap, T., & Levine, S. (2017, May). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In 2017 IEEE international conference on robotics and automation (ICRA) (pp. 3389-3396). IEEE.*

[4] *Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A. J., Banino, A., ... & Hadsell, R. (2016). Learning to navigate in complex environments. arXiv preprint arXiv:1611.03673.*

[5] *Agrawal, P., Carreira, J., & Malik, J. (2015). Learning to see by moving. In Proceedings of the IEEE international conference on computer vision (pp. 37-45).*

[6] *Peng, X. B., Andrychowicz, M., Zaremba, W., & Abbeel, P. (2018, May). Sim-to-real transfer of robotic control with dynamics randomization. In 2018 IEEE international conference on robotics and automation (ICRA) (pp. 3803-3810). IEEE.*

[7] *Florence, P. R., Manuelli, L., & Tedrake, R. (2018). Dense object nets: Learning dense visual object descriptors by and for robotic manipulation. arXiv preprint arXiv:1806.08756.*

[8] *Calandra, R., Owens, A., Jayaraman, D., Lin, J., Yuan, W., Malik, J., ... & Levine, S. (2018). More than a feeling: Learning to grasp and regrasp using vision and touch. IEEE Robotics and Automation Letters, 3(4), 3300-3307.*

[9] *Mirchandani, S., Xia, F., Florence, P., Ichter, B., Driess, D., Arenas, M. G., ... & Zeng, A. (2023). Large language models as general pattern machines. arXiv preprint arXiv:2307.04721.*

[10] *Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., ... & Zitkovich, B. (2022). Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817.*

[11] *Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., ... & Zitkovich, B. (2023). Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818.*

[12] *Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., ... & Zeng, A. (2023, May). Code as policies: Language model programs for embodied control. In 2023 IEEE International Conference on Robotics and Automation (ICRA) (pp. 9493-9500). IEEE.*

[13] *Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., ... & Zeng, A. (2022). Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691.*

[14] *Gupta, A., Savarese, S., Ganguli, S., & Fei-Fei, L. (2021). Embodied intelligence via learning and evolution. Nature communications, 12(1), 5721.*

[15] *Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Wahid, A., ... & Florence, P. (2023). Palm-e: An embodied multimodal language model.*