

Applications of Reinforcement Learning on Humanoid Robot Controlling

Qirui Ge *

School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University,
Shanghai, China

* Corresponding Author Email: iamgqrqwq@sjtu.edu.cn

Abstract. Amidst the swift advancement of artificial intelligence and machine learning technologies, humanoid robots are increasingly demonstrating their potential in mimicking human motions, adapting to intricate environments, and executing a wide array of tasks. Reinforcement Learning (RL), as an advanced learning paradigm that optimizes decision-making processes through environmental interaction, has emerged as a pivotal tool in enhancing the capabilities of humanoid robots. In this paper, the researcher delves into a variety of RL techniques and their implementations, spotlighting key accomplishments and addressing the prevailing challenges alongside envisaging future trajectories. Through an in-depth examination of these applications, our aim is to elucidate the transformative potential of RL in the domain of humanoid robotics, paving the way for more adaptive, intelligent, and autonomous systems. This paper posits that, with the deepening of research and technological progress, RL will catalyze breakthroughs in the humanoid robot sector, propelling smart robots towards greater integration within human society.

Keywords: Reinforcement learning; humanoid robot controlling; applications.

1. Introduction

Reinforcement Learning (RL) has emerged as a powerful paradigm in the realm of artificial intelligence, finding applications across a diverse array of domains, including game control, navigation, and robotics. At its core, RL focuses on the process of learning an optimal policy that maximizes cumulative rewards, showcasing remarkable capabilities in addressing complex manipulation tasks. Traditionally, the majority of RL algorithms rely on predefined reward functions to train agents. While this approach has proven effective, it often falls short in capturing the nuanced intricacies of certain tasks. In recent years, innovative methodologies have surfaced, such as Artificial Curiosity, which aim to guide agents through mechanisms of intrinsic motivation, thereby broadening the scope and enhancing the effectiveness of RL.

The impact of RL is particularly pronounced in the field of robotics, where it has facilitated significant advancements in both simulated environments and real-world applications. Among these applications, the control of humanoid robots stands out due to its inherent complexity and the potential for substantial real-world benefits. Humanoid robots, designed to replicate the human body, necessitate sophisticated control systems to execute tasks that range from simple movements to intricate interactions with their surroundings. RL algorithms have been pivotal in these advancements, enabling the development of robots that can learn from their interactions and progressively improve their performance over time.

This paper aims to delve into the applications of RL in humanoid robot control, offering a comprehensive overview of recent research and developments within this dynamic field. This paper will explore a variety of RL techniques and their implementations, highlight significant achievements, and discuss the challenges that lie ahead, as well as future directions for this captivating area of study. By examining these applications, we aspire to illuminate the transformative potential of RL in revolutionizing humanoid robotics, paving the way for systems that are increasingly adaptive, intelligent, and autonomous.

2. Background

RL is a fundamental paradigm in machine learning where an agent learns to make optimal decisions by interacting with its environment. The agent's goal is to learn an optimal policy that maximizes the cumulative rewards received through these interactions. RL has found widespread applications, particularly in the robotics industry, where it has proven to be a powerful tool for enabling autonomous decision-making and control. Using a value function, Q-Learning assesses the predicted reward for every state-action pair [1]. It functions effectively in discrete action spaces and is straightforward but effective. To handle high-dimensional state spaces, Deep Q-Network blends Q-Learning with deep neural networks. Training is stabilized by innovations like experience replay and fixed target networks [2].

To produce a probability distribution over actions, Policy Gradient directly optimizes the policy function [3]. Complex control problems and continuous action spaces are good fits for it. An enhanced policy gradient technique called Proximal Policy Optimization (PPO) limits the size of policy updates, stabilizing training [4]. It is extensively utilized in robotic control and has demonstrated outstanding performance in challenging conditions. By ensuring that policy updates remain inside a trusted region delineated by the Kullback-Leibler divergence (KL-divergence), Trust Region Policy Optimization (TRPO) guards against significant, unstable changes [5]. Compared to conventional policy gradient techniques, it provides considerable gains in performance and stability.

In order to predict future states and optimize the control approach, Model Predictive Control (MPC) makes use of an environment model [6]. Although it makes effective use of data, the model's accuracy is crucial. A hybrid of Q learning and policy gradient techniques designed for continuous action spaces is called Deep Deterministic Policy Gradient (DDPG) [7]. It makes use of a critic network (value function) and an actor network (policy).

Asynchronous Advantage Actor-Critic (A3C) enhances convergence speed and performance by stabilizing training while concurrently learning the policy and value function through asynchronous multi-threaded updates [8]. An entropy-maximizing reinforcement learning technique called Soft Actor-Critic (SAC) promotes robustness and exploration in policy learning [9]. In challenges involving continuous action space in high dimensions, it exhibits remarkable performance.

3. Application of RL to Humanoid Robotics

In this section, this paper explores the latest advancements in the application of RL to humanoid robotics. The discussion is organized into three key areas: algorithmic optimization, specific scenario applications, and human-like motion imitation.

3.1. Algorithmic Optimization and Development

In the realm of humanoid robotics, the optimization and development of reinforcement learning algorithms play a crucial role in enhancing the performance and robustness of robotic behaviors. This section delves into recent advancements in algorithmic innovations, including the integration of quality diversity with PPO and the development of task space learning methods. These innovations aim to improve the adaptability and efficiency of humanoid robots, enabling them to perform a wide range of tasks with greater precision and flexibility.

Eric et al. [10] introduces a method to enhance reinforcement learning by integrating multiple motion styles into a single policy, called Multi-AMP (Multiple Adversarial Motion Priors) algorithm. The approach builds upon the existing Adversarial Motion Priors (AMP) framework, enabling the policy to switch between various predefined motion styles during training. The method combines task-specific rewards with style rewards derived from discriminators trained to distinguish between the agent's behavior and motion data from different sources. The discriminators help guide the policy to replicate specific styles while performing the required tasks, such as walking or ducking. Multi-AMP is validated on a wheeled-legged quadruped robot, demonstrating advanced skills like transitioning between quadruped and humanoid forms. The experimental results show that Multi-

AMP effectively allows the robot to perform complex tasks by switching between different motion styles without significant performance loss compared to single-task learning.

Guillermo et al. [11] presents a hierarchical framework that integrates RL with model-based control to achieve robust bipedal locomotion. The framework uses a high-level (HL) planner, trained via RL, to generate task space commands based on insights from the Angular Momentum-based Linear Inverted Pendulum (ALIP) model (Fig.1). This HL policy is decoupled from the low-level (LL) controller, which is responsible for tracking the generated trajectories using model-based techniques like inverse dynamics and feedback linearization. The design emphasizes simplicity and efficiency, using a reduced-order state space that captures essential dynamics while simplifying the learning process. The system is tested on multiple bipedal robots, including both underactuated (Rabbit) and fully actuated (Walker2D, Digit) models, demonstrating robust and human-like walking behaviors even under challenging conditions such as varying terrains and external disturbances.

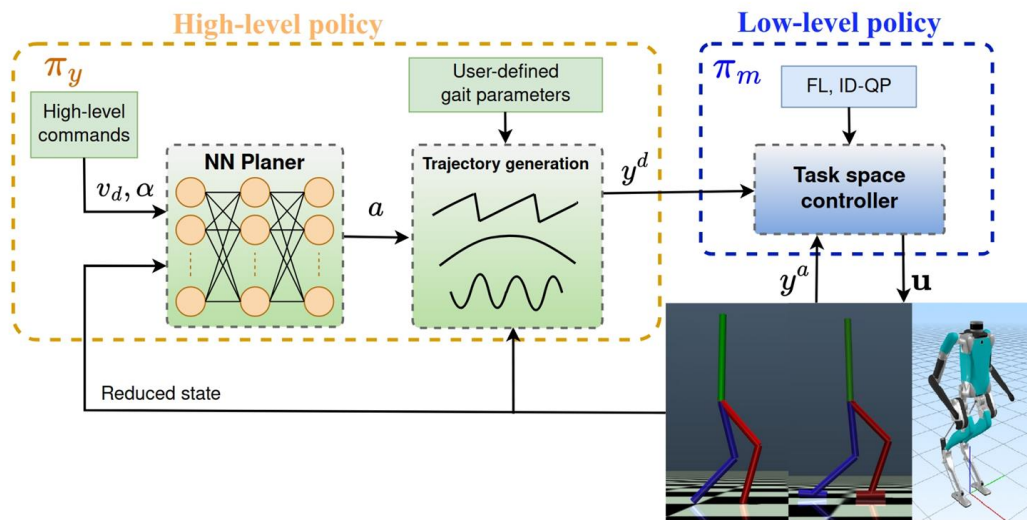


Fig 1. The structure of learning-based framework [11].

The PPGA (Proximal Policy Gradient Arborecence) algorithm is newly introduced by Sumeet et al. [12], which integrates Proximal Policy Optimization (PPO) with Differentiable Quality Diversity (DQD) to enhance the discovery and optimization of diverse, high-performing behaviors in reinforcement learning tasks. The key innovation of this method lies in combining the exploration capabilities of Quality Diversity (QD) algorithms with the robust performance of on-policy reinforcement learning, specifically PPO. The approach allows the agent to explore a wide range of behaviors and optimize them in complex, highdimensional environments like humanoid locomotion tasks. PPGA employs a novel technique of "walking" the search policy across the behavior space, guided by the gradient estimates from PPO, which helps in discovering new and diverse behaviors. This method is validated across multiple continuous control tasks, showing significant improvements over existing QD-RL methods, particularly in challenging domains like humanoid control.

3.2. Learning in Scenarios

Applying reinforcement learning algorithms to real-world scenarios is vital for demonstrating their practical effectiveness. This section explores how these algorithms have been successfully implemented in specific contexts, such as dynamic soccer playing, planned bipedal walking, and the transition from simulation to real-world environments. By focusing on practical applications, these studies illustrate the tangible benefits of reinforcement learning in enhancing the operational capabilities of humanoid robots in complex and varied settings.

Rohan et al. [13] employs a deep RL approach to develop robust bipedal locomotion policies for humanoid robots (Fig.2). Their study focuses on learning a policy capable of following a given step sequence to achieve omnidirectional walking, turning, standing, and stair climbing. The policy is trained using procedurally generated footstep plans and employs a curriculum learning approach to

gradually increase the complexity of the terrains. The RL policy operates within a hierarchical control structure, with a higher-level policy making joint position predictions and a lower-level PD controller executing these predictions. The training is conducted in the MuJoCo simulation environment, using HRP-5P and JVRC-1 humanoid robots. The reward function is carefully designed to ensure that the robot maintains a stable and energy efficient gait while following the planned footsteps. The method's effectiveness is demonstrated through extensive simulation tests, including robustness evaluations under noisy state measurements.

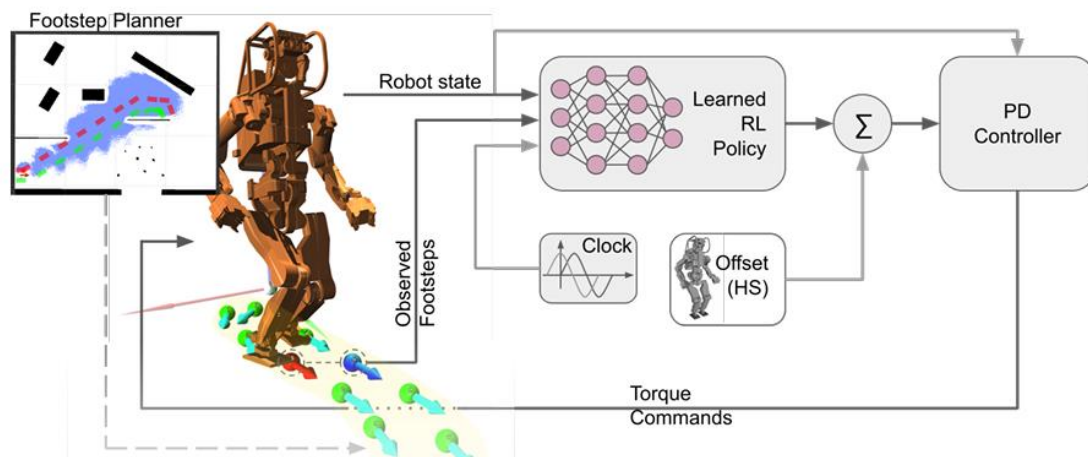


Fig 2. Hierarchical control structure [13].

On the other hand, Rohan et al. [14] presents another method to overcome the sim2real gap in humanoid robots by incorporating current feedback from actuators. Their research focuses on training a deep RL policy in a simulation environment for bipedal locomotion, specifically for the HRP-5P humanoid robot. The key innovation is the simulation of degraded torque-tracking during training, combined with the inclusion of real-time current feedback in the observation space of the RL policy. This approach compensates for the discrepancies between the desired and actual torques applied by the motors due to factors like back electromotive force (EMF). The policy is trained using Proximal Policy Optimization (PPO) in the MuJoCo simulator and then deployed on the real HRP-5P robot, achieving successful bipedal walking, turning, and standing. The effectiveness of the approach is validated through comparisons with conventional model-based controllers, showing superior robustness on uneven terrain.

Ilija et al. [15] also presents a novel learning-based approach for humanoid robot locomotion using a causal transformer model. This model is trained to predict the next action based on the history of proprioceptive observations and actions. The training is performed using large-scale reinforcement learning on thousands of randomized environments in simulation, which is then deployed to the real world without further fine-tuning (zero-shot transfer). The model's ability to adapt in context is a key feature, allowing the robot to navigate various terrains, handle external disturbances, and perform tasks like carrying payloads. The transformer-based approach is validated against state-of-the-art controllers, showing superior performance, especially in adapting to new scenarios and terrains. The policy is designed to handle real-world uncertainties by integrating domain randomization during training, making the sim-to-real transfer effective.

More on to the real-world applications, Tuomas et al. [16] explores the application of deep RL to train a low-cost, miniature humanoid robot to play a simplified one-versus-one soccer game. The research focuses on synthesizing sophisticated and agile movement skills, such as walking, turning, kicking, and rapid fall recovery, which are seamlessly combined to perform complex behavioral strategies in dynamic environments. The agent was trained in a simulated soccer environment using the MuJoCo physics engine, with a two-stage training pipeline: the first stage involved training separate skill policies for tasks like getting up from the ground and scoring goals, while the second stage distilled these skills into a full 1v1 soccerplaying agent through self-play against partially trained copies of itself. Domain randomization and perturbations were applied during training to

ensure the policies were robust enough for zero-shot transfer to real robots. The resulting policy was successfully deployed on real robots, showing superior performance compared to scripted baseline controllers.

3.3. Human Motion Imitation

A key challenge in humanoid robotics is enabling robots to not only perform tasks but to do so in a manner that closely mimics human behavior. This section examines approaches that combine reinforcement learning with imitation learning to achieve human-like motion and expressive control. These researches allow humanoid robots to execute smooth, naturalistic movements and transitions, enhancing their ability to interact with humans and adapt to dynamic environments with a high degree of expressiveness.

Annan et al. [17] introduces a novel approach to enable humanoid robots to replicate human-like locomotion and transition seamlessly between different movement patterns (Fig.3). The research method centers on the use of a Wasserstein adversarial imitation learning system, which leverages a unified primitive skeleton motion retargeting technique. This approach helps to overcome the morphological differences between human demonstrators and humanoid robots, allowing for more natural motion replication. The system incorporates a Wasserstein critic with a soft boundary constraint to stabilize training and prevent mode collapse. The learning process involves conditioning the policy on velocity goals, allowing the humanoid robot to adapt its movements dynamically as speed commands change. The method is tested on the JAXON humanoid robot in a simulated environment, showcasing the robot's ability to perform various complex locomotion behaviors, including walking, running, and push-recovery, with smooth transitions between these actions.

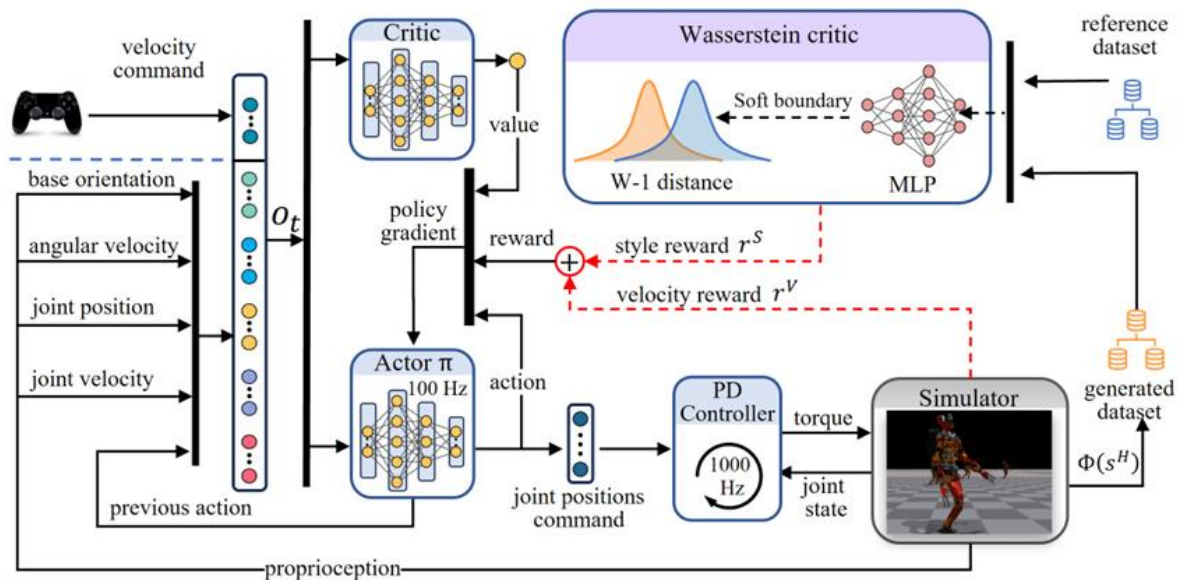


Fig 3. Wasserstein adversarial imitation framework [17].

Also focusing on human imitations, Xuxin et al. [18] presents a method called Expressive Whole-Body Control (ExBody), designed to enable humanoid robots to perform a wide range of expressive motions in realworld environments. The method integrates large-scale human motion capture data with deep RL to train a policy that allows a human-sized robot to mimic diverse human movements. The core idea is to focus on the upper body for expressive motion while relaxing the imitation constraints on the legs, which are instead tasked with following velocity commands robustly. The training is conducted in a highly randomized simulated environment to ensure effective sim-to-real transfer. This approach allows the robot to perform tasks such as dancing, shaking hands, and walking on different terrains, with the ability to adapt its movements dynamically based on commands. The policy is tested on the Unitree H1 humanoid robot, demonstrating robust and naturalistic motions both in simulation and in real-world scenarios.

4. Conclusion

In summary, the application of RL in the domain of humanoid robot control has demonstrated remarkable value. As a revolutionary learning paradigm, RL empowers humanoid robots with unprecedented autonomy and adaptability by continuously refining decision-making through environmental interaction. This paper has explored the latest technological advancements in RL, encompassing algorithm optimization and diversified applications, significantly enriching the functionalities of humanoid robots.

On the algorithmic front, the integration of techniques such as PPO, TRPO, and DDPG has markedly enhanced robotic performance in complex tasks. Moreover, the introduction of innovations like multimodal motion priors and quality diversity has endowed robots with high levels of flexibility and generality. In specific applications, RL has achieved notable successes in humanoid robots, spanning football skills, dynamic environment adaptation, and natural motion imitation, heralding their potential impact across education, healthcare, and daily assistance sectors.

Despite challenges regarding computational resources, training efficiency, and ethical considerations, researchers are actively seeking solutions to ensure that technological advancements are coupled with social responsibility. Future research will concentrate on further optimizing algorithms to enhance learning speed and generalization capacity, while also exploring novel paradigms for human-robot collaboration to foster more harmonious coexistence.

References

- [1] Christopher John Cornish Hellins Watkins and Peter Dayan. Q-learning. In: Machine Learning 8.3 1992, 279–292.
- [2] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with Deep Reinforcement Learning. 2013. arXiv: 1312.5602.
- [3] Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In: Advances in neural information processing systems 12, 1999.
- [4] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. 2017. arXiv: 1707.06347 .
- [5] John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust Region Policy Optimization. 2017. arXiv: 1502.05477 .
- [6] Carlos Eduardo Garcia, David M. Prent, and Manfred Morari. Model predictive control: Theory and practice—A survey. In: Automatica 25.3, 1989, 335–348.
- [7] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. 2019. arXiv: 1509.02971.
- [8] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous Methods for Deep Reinforcement Learning. 2016. arXiv: 1602.01783 [cs.LG].
- [9] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. 2018. arXiv: 1801.01290 [cs.LG].
- [10] Eric Vollenweider, Jose Rojas, Anastasia Mavrommati, and Katja Mombaur. Advanced Skills through Multiple Adversarial Motion Priors in Reinforcement Learning. 2022. arXiv: 2203.14912 [cs.RO].
- [11] Guillermo A. Castillo, Huiyan Chen, Xuxin Cheng, and Zhenglong Sun. Template Model Inspired Task Space Learning for Robust Bipedal Locomotion. 2023. arXiv: 2309.15442 [cs.RO].
- [12] Sumeet Batra, Rohan Pratap Singh, Jack Rae, and Emile Contal. Proximal Policy Gradient Arborecence for Quality Diversity Reinforcement Learning. 2024. arXiv: 2305.13795 [cs.LG].
- [13] Rohan Pratap Singh, Xuxin Cheng, Guillermo A. Castillo, and Zhenglong Sun. Learning Bipedal Walking on Planned Footsteps for Humanoid Robots. 2022. arXiv: 2207.12644 [cs.RO].

- [14] Rohan Pratap Singh, Xuxin Cheng, Guillermo A. Castillo, and Zhenglong Sun. Learning Bipedal Walking for Humanoids with Current Feedback. In: IEEE Access 11 (2023), 82013–82023. issn: 2169-3536.
- [15] Ilija Radosavovic, Rohan Pratap Singh, Xuxin Cheng, and Guillermo A. Castillo. Real-World Humanoid Locomotion with Reinforcement Learning. 2023. arXiv: 2303.03381 [cs.RO].
- [16] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Learning agile soccer skills for a bipedal robot with deep reinforcement learning. In: Science Robotics 9.89 (Apr. 2024). issn: 2470-9476.
- [17] Annan Tang, Xuxin Cheng, Guillermo A. Castillo, and Zhenglong Sun. HumanMimic: Learning Natural Locomotion and Transitions for Humanoid Robot via Wasserstein Adversarial Imitation. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). Vol. 30. IEEE, May 2024, 13107–13114.
- [18] Xuxin Cheng, Guillermo A. Castillo, Rohan Pratap Singh, and Zhenglong Sun. Expressive Whole-Body Control for Humanoid Robots. 2024. arXiv: 2402.16796 [cs.RO].