Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU34599

Motion Control for Quadruped Robots Based on Reinforcement Learning

Xiaotong Wang

Beijing Institute of Technology, School of Mechatronical Engineering, Beijing, China
1120242139@bit.edu.cn

Abstract. Quadruped robots have good locomotive advantages in complex terrains, but traditional control methods rely on precise dynamic modeling and struggle to adapt to unstructured environments. This paper systematically reviews the application of reinforcement learning to quadruped robot motion control and provides references for both academic research and practical application. PPO and Actor-Critic are the two basic algorithms; how they are implemented in training frameworks such as hierarchical control, model predictive control and body perception are explored in this paper. Research has shown that reinforcement learning can enhance the adaptivity and motion stability of robots in complex environments, and good prospects for applications in Mars exploration and disaster relief have been identified. The current main methods still have the problems of low data efficiency, high training costs and difficulty in real-world deployment. Future work will integrate world models, offline learning and lightweight deployment to move these algorithms from the lab to engineering applications.

Keywords: Quadruped robots, reinforcement learning, motion control

1. Introduction

Quadruped robots are relatively mobile and have strong load-bearing capacities. They are much better than other kinds of mobile robots, such as wheeled and tracked robots, in difficult-to-drive environments. Quadruped robots have complex leg structures, and therefore require higher-level motion control. These are the range of joint movement, when to drive the motor, the actual decision-making speed of the algorithm, and how long the motion stays stable. The old way of control is based on dynamic and kinematic modelling; thus, inverse motion control can be achieved through gait scheduling, virtual force control and joint angle determination [1]. Control mode is only for organised areas that have a fixed route. Even with added sensors, it is still not very adaptable to uneven ground and has weak resistance to disturbance.

In recent years, more and more research has been conducted on reinforcement learning for motion control of quadruped robots in complex and unstructured environments to achieve adaptive operations. The core of the decision-making method in Reinforcement Learning algorithms differs significantly from that of traditional control theory. Markov Decision Processes (MDPs) are used to define the state space, action space and reward functions for memory-based modelling. They are not performing static motion planning; they learn in a simulation. They move from virtual trial-and-error learning and convergence of control policy functions to a direct mapping of robot perception to control output [2]. Reinforcement learning algorithms do not need precise dynamic models and their automatically

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU34599

optimised control laws have a relatively short development cycle. Train in a random environment to help quadruped robots adapt to various unstructured terrains and thus extend the application range. They have richer functions and are more flexible; a relatively larger portion of the work can be carried out by machines.

This paper will present some specific training methods for reinforcement learning algorithms in quadruped robot motion control and the properties of different training methods. Summarize the main reinforcement learning algorithms currently employed for quadruped robot motion control, and specify the particular training models that have been applied to or modified by these algorithms. Briefly list some real-world application cases of the trained individual models, and finally summarize which reinforcement learning models are more suitable for which scenarios. It offers suggestions for algorithm optimisation and model application in the future reinforcement learning training of quadruped robots. Combine and compare the designs of quadruped robot reinforcement learning models, simulation training and Sim-to-Real transfer methods proposed by various scholars in this paper. It has conducted research on how model-trained quadruped robots have improved the motion-stability of traditional methods over difficult ground and performed complicated movements.

2. Core implementation algorithms

Reinforcement learning is a way of machine learning that learns to make decisions by interacting with an environment repeatedly. Reinforcement Learning is data-driven and does not use a model. It is generally a Markov Decision Process (MDP), and S and A are the state and action spaces, P is the state transition probability, r is the reward function, and γ is the discount factor. These components constitute the data tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, and are used by the algorithm. The algorithm finds the optimal policy by moving in the direction of the maximum gradient and adjusting hyperparameters on the fly. The state in quadruped robot motion control usually comprises joint encoder data, inertial measurement unit (IMU) information, and historical data. Actions are generally the target joint angles or joint torques. Design a reward function to motivate the behaviour of velocity tracking and posture stability. The purpose of reinforcement learning (RL) is to learn the best policy π^*, that maximizes the cumulative discounted expected reward $E[\sum \gamma^t r_t]$.

Reinforcement learning is model-free or model-light and performs reward-driven automatic parameter optimisation. This means engineers don't have to code machine motion models or spend a lot of time manually tuning parameters. It represents a cutting-edge direction in quadruped robot motion control. Currently, reinforcement learning algorithms applied to quadrupedal robots are mainly divided into policy-based methods (such as Policy Gradient Algorithm PPO) and emerging model-based methods (such as world models and Dyna architecture). In practical applications, reinforcement learning decision-making frameworks are often combined with deep learning information processing methods. The efficient information extraction and encoding capabilities of deep learning help reinforcement learning methods get policy improvement suggestions by processing information more efficiently. As a result, reinforcement learning in machine learning often appears as deep reinforcement learning training methods. In the field of quadruped robot motion control, deep reinforcement learning has two main implementation methods: policy-based Proximal Policy Optimization and model-based Actor-Critic algorithms. The following sections introduce these two core algorithms one by one.

2.1. Proximal Policy Optimization (PPO)

The Proximal Policy Optimization method uses a clipping mechanism or KL penalty to limit the agent's learning rate. This keeps the agent from ending up in a dangerous situation due to steps that

DOI: 10.54254/2755-2721/2026.GU34599

are too large, and also from falling into a too-conservative state due to steps that are too small. Its objective function is as followed:

$$L^{PPO}(\theta) = E_t[\min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t)] \quad (1)$$

where $r_t(\theta) = \pi_\theta(a_t \mid s_t)/\pi_{\theta_{old}}(a_t \mid s_t)$ is the probability ratio of the new and old policies, and $\hat{A}_t$ is the estimated advantage function.

PPO is simple and stable. That is why it has become the most common reinforcement learning method for quadruped robots today. Researchers have used PPO and its variants to get good control performance on quadruped robots. A typical PPO workflow goes like this: first, design a pruning function. Then, interact with the environment to collect data. After that, perform multiple mini-batch updates. After several rounds of curriculum-based training, the final policy is output [3]. To improve a specific performance metric, PPO can be combined with other algorithms. For example, adding Long Short-Term Memory (LSTM) gives PPO a useful memory sequence. This helps the robot learn strategies more efficiently from past experiences [4]. PPO is stable, supports parallel training, and has decent sampling efficiency. These features make it a good fit for continuous motion control of quadruped robots with 8–12 degrees of freedom. But there are also limitations. The small batch updates and the pruning function keep PPO from enabling high-speed running. To achieve fast running, higher sampling frequencies and more carefully designed reward functions are needed.

2.2. Actor-Critic

The Actor-Critic method was proposed to deal with the instability of reinforcement learning in complex, random environments. By training a policy network (Actor) to generate actions and a value network (Critic) to evaluate the quality of the current policy at the same time, it balances exploration and exploitation. This avoids instability caused by learning that is too aggressive or too conservative.

Actor-Critic can get a value function distribution by keeping an ensemble of N Critic networks. It then uses the lower confidence bound of this distribution to evaluate the action, which keeps the value estimate from falling into unrealistic optimism. By using a multi-head Actor-Critic approach, multiple training objectives can be achieved at the same time while maintaining good accuracy. This means quadrupedal robots can maintain good continuity even when performing complex movements [5].

Actor-Critic encourages exploration and allows more efficient policy updates, so it works well for quadrupedal robots operating in complex terrains.

3. Reinforcement learning training methods applied to quadruped robots

Reinforcement learning has the advantages of high efficiency, flexibility, and the ability to execute complex motions. It has become a powerful tool in intelligent machine adaptation and is widely used on quadruped robots to perform movements that traditional control methods cannot do or find difficult to achieve. In practical deployment, the core reinforcement learning algorithms mentioned above are mainly implemented in three ways: hierarchical controllers, methods combining model predictive control (MPC), and body-aware control. These are introduced below.

3.1. Hierarchical controllers

The motion of a quadruped robot is not a single movement but a combination of many basic actions. It is extremely difficult to have the algorithm learn complex movements directly, so hierarchical controller is employed. It allows the robot to autonomously decompose complex movements into individual actions such as running, jumping, and turning, then execute these individual actions in a specific sequence to ultimately fulfill complex motion commands.

DOI: 10.54254/2755-2721/2026.GU34599

A quadruped robot's motion is not a single movement but a combination of many basic actions. Having the algorithm learn complex movements directly is very difficult, so hierarchical controllers are used. They let the robot break down complex movements into individual actions like running, jumping, and turning. Then the robot performs these individual actions in a specific order to finally complete the complex motion command.

When using a hierarchical controller with reinforcement learning, the PPO algorithm is commonly used. Its design, which limits the policy update step size, works well for training two interdependent policies (upper-level and lower-level). In the design of a quadruped robot that can avoid obstacles agilely and transport objects, a hierarchical framework with high-level decision-making and low-level control is used. Through PPO, the robot learns the best action plan for avoiding obstacles. It then breaks this plan into direct low-level control commands sent to the quadruped's actuators to complete the obstacle avoidance and transport tasks [6].

3.2. Reinforcement learning combined with Model Predictive Control (MPC)

Model Predictive Control (MPC) is a relatively traditional method for quadruped robot motion control. It relies on a learnable system model for short-term prediction and single-step control. Since safety constraints in MPC algorithms are written directly as hyperparameters in the code, MPC control provides strong safety protection during the initial model exploration phase. Combining reinforcement learning with MPC makes quadrupedal robot motion control more controllable.

When using reinforcement learning methods combined with MPC, the PPO algorithm is generally the one chosen. Different MPC modeling methods give different control results. For example, state-space modeling with parallel PPO successfully removed tremors in the trotting and sprinting gaits of a quadruped robot [7].

3.3. Body-aware motion

With the integration of deep learning methods and reinforcement learning, quadrupedal robots can now navigate complex terrains without relying on external environmental data for reading, analysis, or perception. Instead, they achieve stable blind gait motion by building up their own movement experience and self-learning.

Reinforcement learning training based on self-awareness reduces reliance on real-world data and speeds up training. This makes it possible for quadrupedal robots to operate in highly variable environments. In this approach, reinforcement learning is mostly implemented using PPO and Actor-Critic methods. The DreamWaQ self-awareness framework uses the PPO reinforcement learning algorithm, and tests on physical robots have shown robust performance in long-distance motion planning [8]. Using the Actor-Critic reinforcement learning algorithm to design a blind gait for quadrupedal robots has allowed them to walk through various unstructured terrains [9].

The integration of deep learning and reinforcement learning represents a new trend in the AI era. With the support of deep reinforcement learning, quadrupedal robots have moved from terrain-based motion design to experience-based motion analysis. This means they will be able to handle more complex natural and man-made environments and find broader application scenarios.

4. Application scenarios

With reinforcement learning, quadruped robots have gained strong adaptability to unstructured terrain, as well as high robustness and stability in motion. They are gradually—or have already—replaced humans in many fields to perform high-risk, highly complex, and highly repetitive tasks, especially in outdoor environments with complex natural conditions. The following section introduces two

DOI: 10.54254/2755-2721/2026.GU34599

emerging application areas for quadruped robots that use reinforcement learning algorithms for motion control.

4.1. Landing and mobility of quadruped robots on asteroids

Asteroid exploration is an inevitable trend in humanity's future development. However, the complex and unfamiliar geographical conditions on asteroids have long been a major challenge that must be overcome. Sending quadruped robots trained with reinforcement learning into space not only saves a lot of money on training and protecting astronauts, but also improves the efficiency of asteroid exploration.

To deal with the challenge of stable navigation caused by the weak surface gravity on asteroids, a quadruped robot using the PPO reinforcement learning algorithm uses randomized transfer learning to improve the model's adaptability to unfamiliar terrain. This quadruped robot successfully performed reorientation and soft landing maneuvers in a simulated environment with irregular, weak gravity [10].

4.2. Complex terrain perception for disaster relief quadruped robots

Post-disaster rescue operations—including rescuing the injured and finding and evacuating trapped people—are extremely important. Reinforcement learning-trained quadruped robots are a good choice for disaster relief missions. In recent years, as reinforcement learning-based motion control for quadruped robots has continued to improve, using such robots for disaster relief and rescue has become feasible in laboratory settings.

A quadruped robot trained with PPO has navigation and dexterous movement abilities. In laboratory tests, it can accurately perform complex movements like climbing stairs and going around curves [11]. It can take part in initial search operations during the early stages of a fire. Disaster response robots entering the fire scene could quickly find victims and flammable materials early on, giving rescue personnel useful information for planning search and rescue routes and assigning rescue tasks.

5. Conclusion

This paper reviewed the mainstream algorithmic implementations of reinforcement learning in quadruped robot motion control. It then introduced the practical modeling approaches for the three main quadruped robot control models that incorporate reinforcement learning. Finally, two relatively new application scenarios were presented that have shown clear advantages for reinforcement learning-based quadruped robots. Based on the above analysis, it can be concluded that quadruped robots controlled by reinforcement learning algorithms currently demonstrate excellent performance in adaptability, robustness and task completion efficiency in unstructured environments. They are a new stage of development for embodied intelligence in general.

However, most of the research results in this area are still confined to individual technical achievements within the laboratory. At present, few examples of reinforcement learning-based quadruped robots have been widely applied in actual production. No single reinforcement learning-based quadruped robot has achieved a good balance of high efficiency, safety, strong adaptability and low cost to date. In addition, the current algorithms (such as PPO) are required to have a large amount of interaction data and a long training period; therefore, they are difficult to improve quickly. The potential of high sample efficiency for Actor-Critic and DQN methods has yet to be fully realised.

In the future, reinforcement learning training for quadrupedal robots will likely incorporate world models or experience replay enhancement mechanisms to reduce training costs. A small model can be used to reduce cost and move reinforcement learning-based quadrupedal robots out of the lab and

DOI: 10.54254/2755-2721/2026.GU34599

into real-world applications. DQN-class methods can be enhanced by combining them with the new field of offline reinforcement learning. Although quadruped robots have been greatly enhanced by reinforcement learning algorithms, there are still several practical problems in their real-world applications, such as safety, efficiency and generalization; therefore, more research needs to be carried out to address these deficiencies and advance the development of reinforcement learning methods.

References

[1] C. Gehring, S. Coros, M. Hutter, M. Bloesch, M.A. Hoepflinger, R. Siegwart, Control of dynamic gaits for a quadrupedal robot, in Proceedings of the 2013 IEEE International Conference on Robotics and Automation (ICRA), Karlsruhe, Germany, May 6-10 (2013), 3287–3292. https://doi.org/10.1109/ICRA.2013.6631035

[2] M. Aractingi, P.A. Léziart, T. Flayols, Controlling the Solo12 quadruped robot with deep reinforcement learning. Sci. Rep. 13, 11945 (2023). https://doi.org/10.1038/s41598-023-38259-7

[3] P. Escudero-Villa, G.D. Machado-Merino, J. Paredes-Fierro, Quadruped robot locomotion based on deep learning rules. Eng. Proc. 87, 100 (2025).

[4] Y. Qi, B. Yu, S. Xu, Application of autonomous control for quadruped robots based on deep reinforcement learning, in Proceedings of the 6th International Conference on Communications, Information System and Computer Engineering (CISCE) (2024), 1299–1302.

[5] Y. Fan, Z. Pei, H. Shi, Quadruped robot locomotion via soft actor-critic with multi-head critic and dynamic policy gradient. Appl. Intell. 55, 720 (2025).

[6] D. Azimi, R. Hoseinnezhad, Hierarchical reinforcement learning for quadrupedal robots: efficient object manipulation in constrained environments. Sensors 25, 1565 (2025). https://doi.org/10.3390/s25051565

[7] M.R.H. Yazdi, M. Pourghavam, R. Bidari, Hybrid control of advanced quadruped locomotion: integrating model predictive control with deep reinforcement learning, in Proceedings of the 12th RSI International Conference on Robotics and Mechatronics (ICRoM) (2024), 201–208.

[8] I.M.A. Nahrendra, B. Yu, H. Myung, Dreamwaq: learning robust quadrupedal locomotion with implicit terrain imagination via deep reinforcement learning, in Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA) (2023), 5078–5084.

[9] S.D. Marcus, A. Shapiro, C. Giladi, Enhancing quadruped robot walking on unstructured terrains: a combination of stable blind gait and deep reinforcement learning. Electronics 14, 1431 (2025).

[10] J. Qi, H. Gao, H. Su, Reinforcement learning and sim-to-real transfer of reorientation and landing control for quadruped robots on asteroids. IEEE Trans. Ind. Electron. 71, 14392–14400 (2024).

[11] B. Huang, B. Huang, Y. Hou, Training and simulation of quadrupedal robot in adaptive stair climbing for indoor firefighting: an end-to-end reinforcement learning approach, arXiv preprint (2026). https://doi.org/10.48550/arXiv.2602.03087