31st International Conference
ENGINEERING MECHANICS 2025
Medlov, Czech Republic, May 12 –14, 2025
Paper **#3,** pp. 209–212, doi: 10.21495/em2025-209

STABLE WALKING FOR BIPEDAL ROBOTS USING PPO

Vechet S.[1], **Krejsa J.**[2], **Chen K.S.**[3]

Abstract: *This paper explores the application of Proximal Policy Optimization (PPO) for achieving stable bipedal locomotion in autonomous robots. Unlike traditional model-based approaches, PPO leverages reinforcement learning to develop adaptive gait patterns that respond to environmental changes in real time.*

Keywords: Machine Learning, Proximal Policy Optimization, Walking Robot, Biped Robot.

1. Introduction

Autonomous bipedal robots have gained significant attention in recent years, driven by advancements in control algorithms, reinforcement learning, and robotic actuation. Unlike wheeled robots (Thrun et al., 2005; Vechet et al., 2020), which benefit from inherent stability, bipedal robots must continuously adjust their posture and gait to maintain balance while navigating complex environments (Vechet, 2011; Krejsa and Vechet, 2018). This challenge is particularly relevant in applications ranging from industrial automation to assistive robotics and disaster response, where humanoid robots must operate in unstructured terrains. More recent advancements in model-based control (Vechet et al., 2024), such as Model Predictive Control (MPC), improved stability by incorporating real-time feedback and optimizing foot placement strategies. However, these approaches often require precise modeling of the robot's dynamics and are computationally expensive, limiting their real-time applicability in highly dynamic scenarios.

In contrast, reinforcement learning (RL) has emerged as a promising alternative for bipedal locomotion control, offering a data-driven approach to learning complex motor behaviors. Among RL algorithms, Proximal Policy Optimization (PPO) has shown remarkable success in training robust policies for continuous control tasks (Schulman et al., 2017). PPO leverages a trust region-based update mechanism to ensure stable and efficient policy learning, making it well-suited for high-dimensional problems such as bipedal walking. By optimizing a reward function that accounts for stability, energy efficiency, and step regularity, PPO-based controllers can generate adaptive gait patterns that respond to environmental changes in real time.

Despite the potential of PPO for bipedal locomotion, several challenges remain, including sample efficiency, sim-to-real transfer, and robustness to external perturbations (Grandia et al., 2024). We present an approach to training bipedal locomotion policies to improve stability and through experiments, we demonstrate that PPO-based control strategies can achieve stable and adaptive walking behaviors in real-world robot.

2. Materials and Methods

This section describes the tools and methodologies used to develop a simulation model of a bipedal walking robot PAWO (see Fig. 1) and implement Proximal Policy Optimization (PPO) for stable locomotion. The simulation framework is built using PyBullet (Vechet et al., 2024, 2020), a physics engine that enables accurate modeling of robotic movements and interactions with the environment.

[1] Assoc. Prof. Stanislav Vechet, Ph.D.: Institute of Thermomechanics AS CR, v.v.i., Technicka 2, 616 69, Brno; CZ, vechet@it.cas.cz

[2] Assoc. Prof. Jiri Krejsa, Ph.D.: Institute of Thermomechanics AS CR, v.v.i., Technicka 2896/2, 616 69, Brno; CZ, krejsa@it.cas.cz

[3] Prof. Kuo-Shen Chen, PhD.: National Cheng Kung University, Department of Mechanical Engineering, No.1, Ta-Hsueh Road, Tainan 701; Taiwan, kschen@ncku.edu.tw

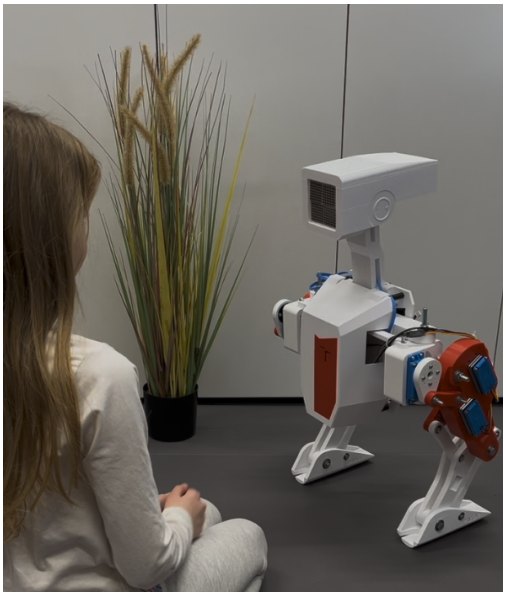

Fig. 1: PAWO Personal Autonomous Walking Optimizer robot in it's latest design configuration during social interaction with human operator.

PyBullet is a Python-based physics simulation library built on Bullet Physics, a widely used open-source engine for real-time collision detection and multi-body dynamics. It provides an efficient platform for simulating robotic systems, allowing precise control over physics parameters such as gravity, joint forces, and contact dynamics.

The bipedal robot is modeled using the Unified Robot Description Format (URDF) to define its mechanical structure, including links, joints, and actuation properties. The robot, consisting of ten rigid bodies connected by joints, is imported into PyBullet, creating a digital twin that replicates its physical behavior (Vechet et al., 2024). The simulation environment is configured with realistic physics parameters, including gravitational acceleration and motor constraints, to match real-world conditions.

Proximal Policy Optimization (PPO) is implemented to learn stable walking behaviors by continuously adjusting joint torques based on sensory feedback. The control policy is trained in simulation using reward functions that encourage balance, energy efficiency, and smooth gait transitions. Position control is used in the simulation, aligning with the actuation method of the real robot's servo-driven joints.

2.1. PPO Algorithm

The PPO algorithm optimizes the policy by iteratively updating parameters while ensuring that changes remain within a safe trust region. The objective function for PPO is given by:

$$L(\theta) = \mathbb{E}\left[\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t)\right] \tag{1}$$

where:

- $r_t = r_t^{move\_forward} + r_t^{falling\_forward} + r_t^{rotate}$ is reward function taking into account translation of center of the gravity, falling forward (pitch), rotation around vertical axis (yaw),
- $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is the probability ratio between the new and old policies,
- A_t is the advantage function,
- ϵ is a small clipping parameter to prevent excessive policy updates (see Table 1).

The PPO training procedure follows these steps:

Algorithm 1 Proximal Policy Optimization (PPO)

1: Initialize policy parameters θ and value function parameters ϕ
2: **for** each iteration **do**
3: Collect trajectories using policy π_θ
4: Compute advantage estimates A_t
5: Optimize surrogate loss $L(\theta)$ using clipped objective
6: Update value function by minimizing squared error loss
7: Update policy using gradient ascent
8: **end for**

The policy is updated using stochastic gradient descent based on the clipped objective function. PPO ensures stable learning by preventing large policy updates, making it suitable for bipedal locomotion tasks where stability is critical.

2.2. Simulation Parameters

Table 1 summarizes the key parameters used in the simulation:

Parameter	Value
Simulation time step	$1/240$ s
Gravity	-9.81 m/s^2
PPO learning rate	3×10^{-4}
Clip parameter ϵ	0.2
Discount factor γ	0.99
Reward function weights	
Move forward w_{fwd}	0.9
Falling forward w_{pitch}	0.9
Rotate around vertical axis w_{yaw}	0.05

Tab. 1: Simulation and PPO training parameters.

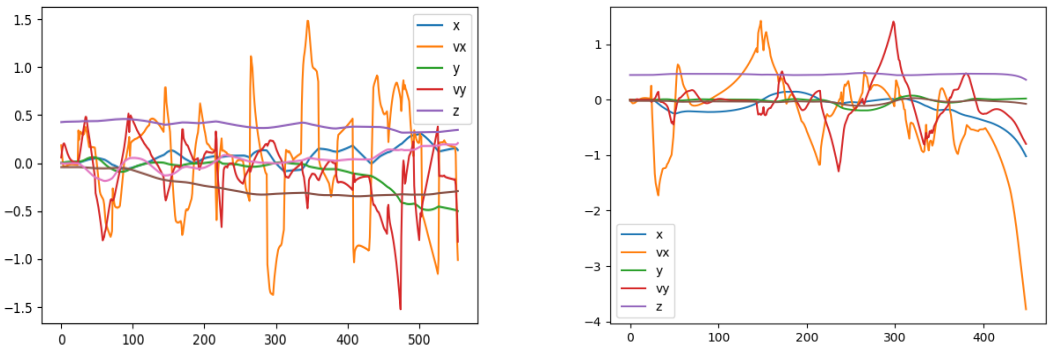

Fig. 2: Telemetry data gathered during walking after training (successful left, unsuccessful right) in the distance traveled on the horizontal axis in millimeters. Both figures shows $x, y, z[mm]$ coordinates of center of the gravity for robots torso and angular velocity for axis $x, y[rad.s^{-1}]$. The rotation around vertical z axis is not shown. The walking attempt on the right was canceled during exceeding maximal value for angular velocity in the x axis which means the robot is falling front.

3. Conclusions

The successful application of Proximal Policy Optimization (PPO) to the dynamic walking of an underpowered and underactuated bipedal robot highlights the potential of reinforcement learning in optimizing

locomotion strategies for constrained robotic systems. The observed variations in episode lengths during training align with expected reinforcement learning behaviors, including the exploration-exploitation trade-off, policy stability mechanisms, and dynamic interactions with the environment.

This work demonstrates that PPO can effectively enable stable and adaptive walking in connection with real world applications where unexpected disturbances can be observed. The learned policies exhibit robustness to varying environmental conditions and perturbations, making this approach a viable solution for real-world robotic locomotion challenges.

Looking ahead, these findings pave the way for applying similar methodologies to larger and heavier humanoid robots with industrial applications. Scaling up PPO-based control frameworks will require addressing additional challenges, such as increased inertia, higher degrees of freedom, and the need for real-time adaptation in unstructured environments. Future research will explore the integration of sim-to-real transfer techniques, domain adaptation strategies, and hardware-in-the-loop training to bridge the gap between simulated and physical deployments.

As reinforcement learning continues to mature, its application in humanoid robotics is expected to advance the field significantly, contributing to the development of agile, efficient, and autonomous bipedal robots for industrial and commercial use.

Acknowledgement

This study was realized with the institutional support RVO: 61388998.

References

Grandia, R., Knoop, E., Hopkins, M., Wiedebach, G., Bishop, J., Pickles, S., Müller, D., and Bächer, M. (2024) Design and control of a bipedal robotic character. In *Robotics: Science and Systems XX*. DOI: 10.15607/RSS.2024.XX.103.

Krejsa, J. and Vechet, S. (2018) Fusion of local and global sensory information in mobile robot outdoor localization task. In Maga, D., Stefek, A., and Brezina, T., eds, *Proceedings of the 2018 18th International Conference on Mechatronics - Mechatronika (ME)*. pp. 296–300.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017) Proximal policy optimization algorithms. *arXiv*. https://arxiv.org/abs/1707.06347.

Thrun, S., Burgard, W., and Fox, D. (2005) *Probabilistic Robotics*. MIT Press.

Vechet, S. (2011) The rule based path planner for autonomous mobile robot. In Matousek, R., edt., *MENDEL 2011 - 17th International Conference on Soft Computing*. B&R Automat CZ Ltd; Humusoft Ltd; AutoCont CZ Ltd, pp. 546–551.

Vechet, S., Krejsa, J., and Chen, K.-S. (2020) AGVs mission control support in smart factories by decision networks. In Maga, D. and Hájek, J., eds, *Proceedings of the 2020 19th International Conference on Mechatronics - Mechatronika (ME)*. pp. 1–4. DOI: 10.1109/ME49197.2020.9286465.

Vechet, S., Krejsa, J., and Chen, K.-S. (2024) Vertical stabilization of bipedal walking drone PAVO with proximal policy optimization. In *Proceedings of the 2024 21st International Conference on Mechatronics - Mechatronika (ME)*. pp. 1–6. DOI: 10.1109/ME61309.2024.10789752.