NDSEG

NATIONAL DEFENSE SCIENCE AND ENGINEERING GRADUATE FELLOWSHIP PROGRAM

Machine Learning for Quadruped Robots

Towards Learning High-Value Footstep Placements

Electrical Engineering

Jeremiah M Coholich

Georgia Institute of Technology

Fellowship Year: 2020

Service: ONR

Advisor/Mentor: Dr. Zsolt Kira

INTRODUCTION

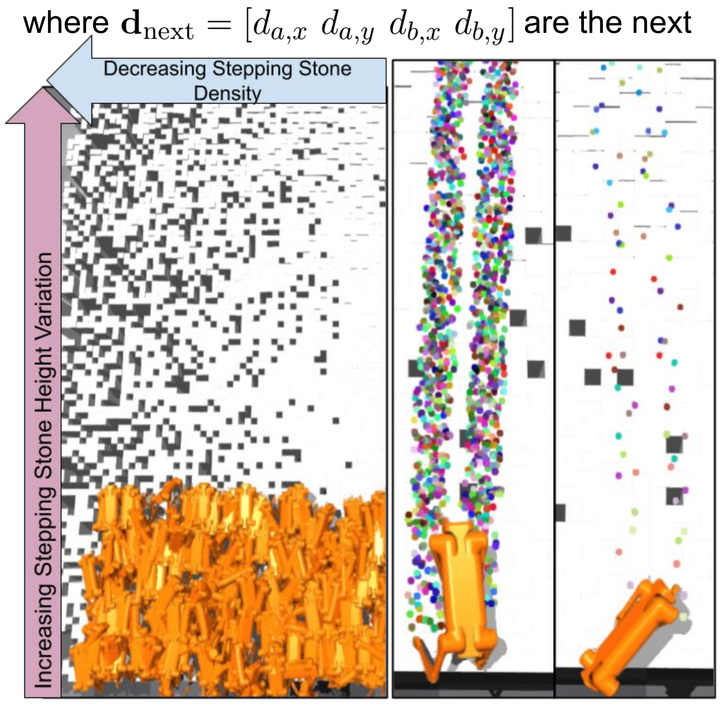
Legged robots are useful platforms for traversing treacherous terrain. Recently, there has been an explosion of interest in using reinforcement learning (RL) for the control of quadruped robots [1][2]. However, the naive application of RL often results in policies that exhibit strange motions that are unsuitable for real-world deployment. In this work, we hope to overcome these issues with a novel RL-based footstep planner and joint-level controller for traversing stepping stones. By providing an intermediate goal of hitting footstep targets, we hope to learn realistic and successful locomotion policies.

METHODS

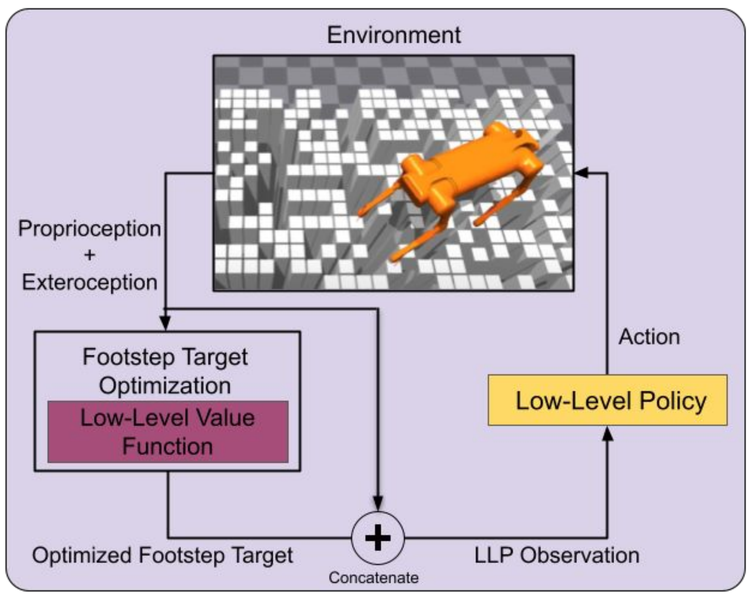
First, we train a low-level policy (LLP) to hit randomly-generated footstep targets over stepping stone terrain (bottom figure) using the Proximal Policy Optimization (PPO) [3] algorithm and a reward function that encourages the robot to hit the footstep targets. During the course of training, we learn a value function (below) which estimates the discounted sum of future rewards at the current state $\mathbf{s}_0$ given the current policy π. The state of the robot includes information about the robot's joints, the surrounding terrain, and the location of the next footstep targets.

$$V_\pi(\mathbf{s_0}) := \mathbb{E}_\pi\left[r(\mathbf{s_0}) + \sum_{t=1}^{H} \gamma^t r(\mathbf{s}_t)\right]$$

Next, we define a high-level policy (HLP) which selects optimal footstep targets with respect to the value function of the LLP and an objective that rewards selecting targets in a desired direction. The HLP solves the optimization problem (next column), where $\mathbf{d}_{\text{next}} = [d_{a,x}\ d_{a,y}\ d_{b,x}\ d_{b,y}]$ are the next



The training environment for the Low-Level Policy. Left: Thousands of robots are simulated in parallel on terrain with varying difficulty. Right: Two randomly generated sequences of footstep targets.



Our proposed planning and control architecture. The Low-Level Policy is learned via reinforcement learning. The footstep target optimization is defined in the equations below.

footstep targets, α is the desired robot heading, R is a 3D rotation matrix, and k is a hyperparameter that controls the tradeoff between the value-based term and the directional term.

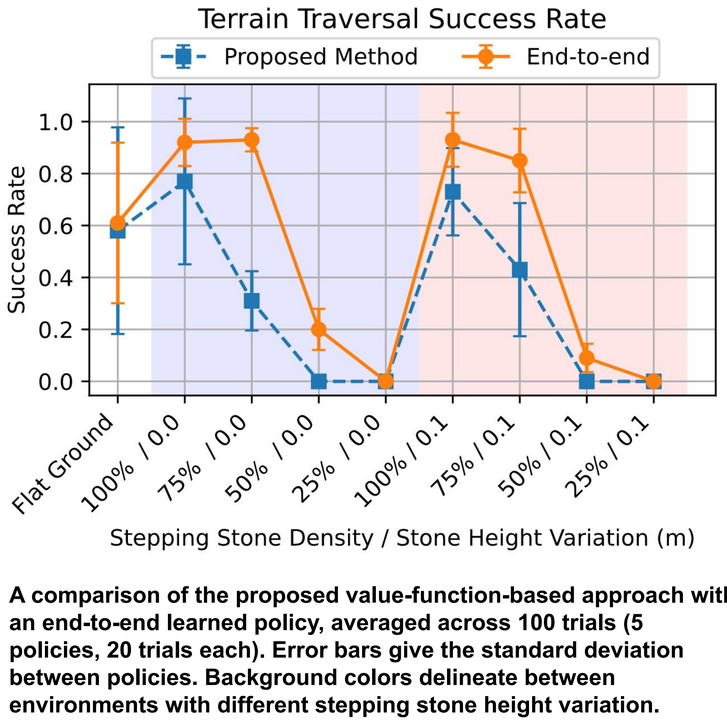
$$\mathbf{H} = \begin{bmatrix}\cos\alpha & \sin\alpha & 0\end{bmatrix} R_z(\theta_{yaw}) \begin{bmatrix} d_{a,x} & d_{b,x} \\ d_{a,y} & d_{a,y} \\ 0 & 0 \end{bmatrix}$$

$$\mathbf{d}^*_{\text{next}} = \arg\max_{\mathbf{d}_{\text{next}}} V(\mathbf{s}_0) + k\mathbf{H}$$

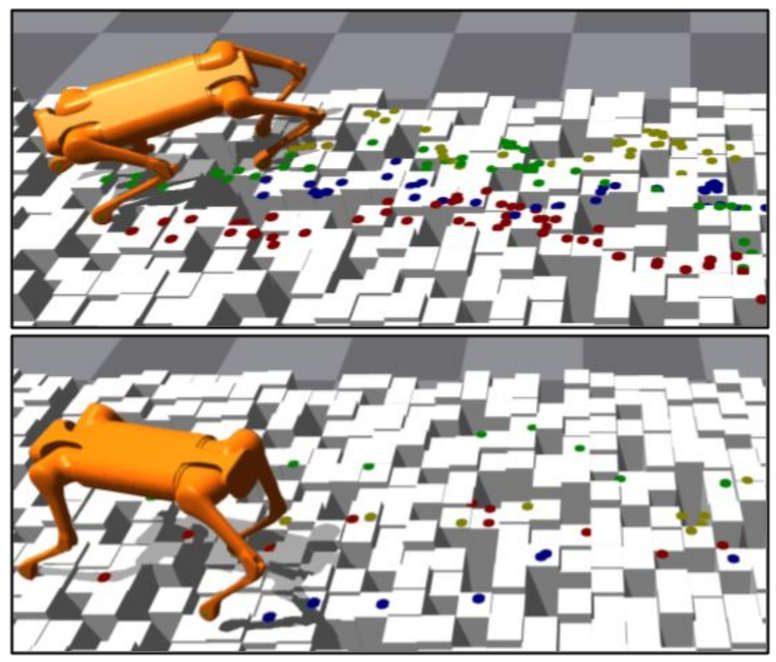
We compare this approach to an end-to-end policy trained with reinforcement learning, where the robot is simply rewarded for its velocity in the forward direction over the stepping stone terrain.

RESULTS

To evaluate the performance of our method, we test the quadruped's ability to traverse stepping stone terrains of varying difficulty. We vary the density of the stepping stones (higher density is easier) and the random height variation of the stepping stones (lower is easier). Success is defined as traversing an 8.0 m



A comparison of the proposed value-function-based approach with an end-to-end learned policy, averaged across 100 trials (5 policies, 20 trials each). Error bars give the standard deviation between policies. Background colors delineate between environments with different stepping stone height variation.



Quadruped crossing the stepping stones terrain with density of 75% and 0.1 meter height variation using our proposed approach (top) and an end-to-end learned policy (bottom). The previous foot contact locations with the terrain are plotted.

section of terrain without falling or colliding with anything. Our method achieves some success in traversing difficult terrain, but does not outperform the end-to-end approach.

CONCLUSION

Our method is able to cross stepping stone terrain, but its performance can be improved. In the future, we plan to experiment with the following:

- Applying temporal smoothing to the HLP
- Reducing the differences between the LLP training environment and the evaluation environments
- Replacing the "greedy" directional term in the objective function of the HLP

BENEFITS TO DOD

Quadruped robots are unmanned and highly versatile platforms which can be deployed for missions such as search-and-rescue, reconnaissance, and explosive ordnance disposal [4]. This research targets locomotion over difficult terrain, which is frequently encountered in warzones. The end goal is to use quadrupeds to support or replace military personal in dangerous environments.

REFERENCES

[1]Haarnoja, Tuomas, Sehoon Ha, Aurick Zhou, Jie Tan, George Tucker, and Sergey Levine. "Learning to Walk Via Deep Reinforcement Learning." In Robotics: Science and Systems. 2019.

[2]Lee, Joonho, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. "Learning quadrupedal locomotion over challenging terrain." Science robotics 5, no. 47 (2020): eabc5986.

[3] Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. "Proximal policy optimization algorithms." arXiv preprint arXiv:1707.06347 (2017).

[2] Jorgensen, Steven Jens, Michael W. Lanighan, Sylvain S. Bertrand, Andrew Watson, Joseph S. Altemus, R. Scott Askew, Lyndon Bridgwater et al. "Deploying the nasa valkyrie humanoid for ied response: An initial approach and evaluation summary." In 2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids), pp. 1-8. IEEE, 2019.

Anticipated Graduation: May 2025

www.jeremiahcoholich.com