On-line Optimization of Dynamic Walking Using Stochastic Policy Gradient Ascent

Jörg Stückler and Sven Behnke
Computer Science Institute, University of Freiburg, Germany
{stueckle | behnke}@informatik.uni-freiburg.de, http://www.NimbRo.net

INTRODUCTION

In this abstract, we present a method for on-line optimization of dynamic walking patterns for a humanoid robot. The 19DOF robot has human-like proportions, weighs 2.3kg, and is 60cm tall. A clock-driven gait engine [1] generates joint trajectories for various forward speeds. We use a hierarchy of open-loop and closed-loop control policies, each manipulating an individual subset of the gait engine parameters. These policies are improved by stochastic policy gradient ascent. The nominal weight shifting of the gait at zero speed is characterized by the rotational roll and pitch motion of the upper trunk during the complete cycle. We reward imitation of this nominal weight shifting and measured walking speed. The policy gradients are estimated by the gradients of the long-term expected accumulated reward. The algorithm must handle non-stationarity of the learning problem caused by continuous changes in target speed. It must also cope with the dependencies between open-loop and closed-loop policy learning. Related methods [2-3] have been applied recently to simpler robots.

METHODS

We apply gait-cycle synchronous Fourier analysis to characterize the roll and pitch motion of the upper trunk. We use the first 6 complex coefficients for the roll axis, but only the first coefficient for the attitude pitch. After each gait cycle, the robot is rewarded with the weighted sum of the negative mean difference between actual and example Fourier coefficients. The achieved walking speed is also rewarded.

The open-loop component learns a stochastic policy for all 18 parameters of the gait engine. By estimating the average reward following the current policy, the algorithm performs hill-climbing to improve the policy according to the change in the value estimate. The closed-loop learning component is implemented in an actor/critic reinforcement learning scheme [4]. It controls shifting amplitude and phase shift with a stochastic policy. We use amplitude and phase shift of the second Fourier coefficient of the attitude roll motion as feedback after each gait cycle. The critic estimates the value function with linear function approximation and TD(λ) prediction. The actor manipulates the stochastic policy according to the TD(λ)-error in the critic. Because jumps in trajectories need to be avoided, we adapt the gait parameters smoothly to new values. To handle the resulting delay we use eligibility traces in the actor.

RESULTS AND DISCUSSION

To assess the performance of the algorithm, we let the robot walk with slowly increasing target speed starting at zero speed. Fig.1 and Fig.2 show the estimated value of the open-loop learning component and walking speed, respectively, with and without learning. When learning, the robot can walk at much higher gait target speeds. It also imitates the nominal weight shifting better at lower target speeds, which yields a more stable walking pattern.

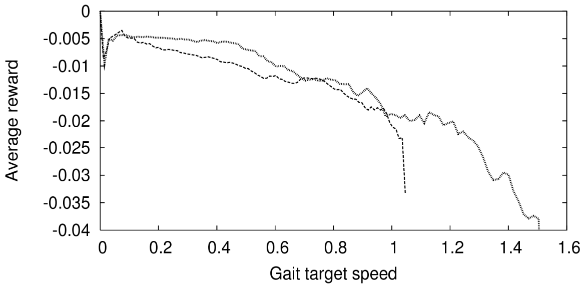

Figure 1 : Average reward as estimated by the open-loop component with learning enabled (dotted) and disabled (dashed)

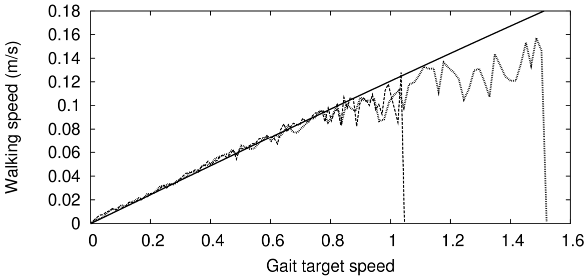

Figure 2 : Achieved walking speed with learning enabled (dotted) and disabled (dashed). The solid line is the nominal walking speed.

CONCLUSIONS

Our algorithm improves walking patterns for bipedal locomotion based on a performance measure that is evaluated in each gait cycle. The measure rewards imitation of nominal weight shifting and walking speed. The combination of open-loop and closed-loop learning resulted in improved walking stability and higher speeds.
Currently, the method is evaluated for the simulated robot. Because the gait engine works well for the physical robot, we expect the results to carry over to the real robot. So far, we use only slow feedback. We plan to integrate a closed-loop component that learns a control policy for a rate of 83Hz to compensate for fast disturbances.

REFERENCES

1. S. Behnke. Online Trajectory Generation for Omnidirectional Biped Walking. In Proc. of ICRA, 05/2006
2. R. Tedrake, T. W. Zhang, and H. S. Seung. Learning to Walk in 20 Minutes. Proc. of the Fourteenth Yale Workshop on Adaptive and Learning Systems, 2005.
3. T. Geng, B. Porr, F. Wörgötter. Fast Biped Walking with a Sensor-driven Neuronal Controller and Real-time Online Learning. J. of Robotics Research, 2006.
4. R. Sutton and S. Barto. Reinforcement Learning: An Introduction. MIT Press, 1998