Poincaré-Map-Based Reinforcement Learning For Biped Walking

Jun Morimoto[1,2], Jun Nakanishi[1,2], Gen Endo[2,3], and Gordon Cheng[1,2]
[1]*Computational Brain Project, ICORP, JST*
[2]*ATR Computational Neuroscience Laboratories*
[3]*Sony Intelligence Dynamics Laboratories, Inc.*
2-2-2 Hikaridai Soraku-gun Seika-cho, Kyoto, 619-0288, JAPAN
xmorimo@atr.jp

Christohper G. Atkeson and Garth Zeglin
The Robotics Institute
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA, 15213, USA
cga@cs.cmu.edu

— We propose a model-based reinforcement learning algorithm for biped walking in which the robot learns to appropriately modulate an observed walking pattern. Via-points are detected from the observed walking trajectories using the minimum jerk criterion. The learning algorithm modulates the via-points as control actions to improve walking trajectories. This decision is based on a learned model of the Poincare´ map of the periodic walking pattern. The model maps from a state in the single support phase and the control actions to a state in the next single support phase. We applied this approach to both a simulated robot model and an actual biped robot. We show that successful walking policies are acquired.

***Index Terms*— Biped Walking; Reinforcement Learning; Poincare´ map**

I. INTRODUCTION

We propose a learning algorithm to acquire an appropriate biped walking controllers by modulating an observed walking pattern. We are using model-based reinforcement learning, where we learn a model of a Poincare´ map and then choose control actions based on a computed value function. We detect via-points from an observed walking trajectory and use the via-points as control actions.

Several researchers have applied reinforcement learning to biped locomotion [12], [2]. Few studies deal with a physical robots because reinforcement learning methods often require large numbers of trials. The policy gradient method [17] is one of the reinforcement learning methods successfully applied to learn biped walking on actual robots [1], [19]. However, [1] requires hours to learn a walking controller, and [19] requires a mechanically stable robot.

On the other hand, [3] reported that a model-based approach to reinforcement learning is able to accomplish given tasks much faster than without using knowledge of the environment. In our previous work [11], we showed that a model-based approach using an approximated Poincare´ map could be applied to learn biped walking in small numbers of trials. However, we used an empirically designed nominal trajectory for the proposed method, and acquired a successful walking pattern only in a simulated environments. In this study, we use observed trajectories, such as those of humans or other robots controlled by this or other algorithms, as nominal trajectories. We show that the proposed method can be applied to an actual robot (Fig. 1).

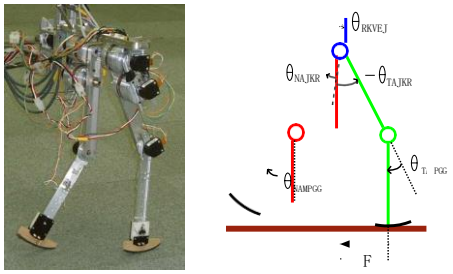


Fig. 1. Five link biped robot. Input state $\mathbf{x} = (d, \dot{d})$

First, we use a simulated 5 link biped robot (Fig. 1) to evaluate our proposed method. Physical parameters of the 5 link simulated robot in *TABLE I* are selected to model the actual biped robot fixed to a boom that keeps the robot in the sagittal plane (Fig. 1). Our biped has a short torso and round feet without ankle joints. For these bipeds, controlling biped walking trajectories with the popular ZMP approach [7], [22] is difficult or impossible, and thus an alternative method for controller design must be used.

In section II, we introduce our reinforcement learning method for biped walking. In section III, we show simulation results. In section IV, we present an implementation of the proposed method on the real robot, and then demonstrate that the robot acquires a successful walking pattern within 100 trials.

II. POINCARE´-MAP-BASED REINFORCEMENT LEARNING FOR BIPED LOCOMOTION

We improve biped walking controllers based on an approximated Poincare´ map using a model-based reinforce-

Poincaré-Map-Based Reinforcement Learning For Biped Walking

Jun Morimoto[1,2], Jun Nakanishi[1,2], Gen Endo[2,3], and Gordon Cheng[1,2]
[1]*Computational Brain Project, ICORP, JST*
[2]*ATR Computational Neuroscience Laboratories*
[3]*Sony Intelligence Dynamics Laboratories, Inc.*
2-2-2 Hikaridai Soraku-gun Seika-cho, Kyoto, 619-0288, JAPAN
xmorimo@atr.jp

Christohper G. Atkeson and Garth Zeglin
The Robotics Institute
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA, 15213, USA
cga@cs.cmu.edu

— We propose a model-based reinforcement learning algorithm for biped walking in which the robot learns to appropriately modulate an observed walking pattern. Via-points are detected from the observed walking trajectories using the minimum jerk criterion. The learning algorithm modulates the via-points as control actions to improve walking trajectories. This decision is based on a learned model of the Poincaré map of the periodic walking pattern. The model maps from a state in the single support phase and the control actions to a state in the next single support phase. We applied this approach to both a simulated robot model and an actual biped robot. We show that successful walking policies are acquired.

***Index Terms*— Biped Walking; Reinforcement Learning; Poincaré map**

I. Introduction

We propose a learning algorithm to acquire an appropriate biped walking controllers by modulating an observed walking pattern. We are using model-based reinforcement learning, where we learn a model of a Poincaré map and then choose control actions based on a computed value function. We detect via-points from an observed walking trajectory and use the via-points as control actions.

Several researchers have applied reinforcement learning to biped locomotion [12], [2]. Few studies deal with a physical robots because reinforcement learning methods often require large numbers of trials. The policy gradient method [17] is one of the reinforcement learning methods successfully applied to learn biped walking on actual robots [1], [19]. However, [1] requires hours to learn a walking controller, and [19] requires a mechanically stable robot.

On the other hand, [3] reported that a model-based approach to reinforcement learning is able to accomplish given tasks much faster than without using knowledge of the environment. In our previous work [11], we showed that a model-based approach using an approximated Poincaré map could be applied to learn biped walking in small numbers of trials. However, we used an empirically designed nominal trajectory for the proposed method, and acquired a successful walking pattern only in a simulated environments. In this study, we use observed trajectories, such as those of humans or other robots controlled by this or other algorithms, as nominal trajectories. We show that the proposed method can be applied to an actual robot (Fig. 1).

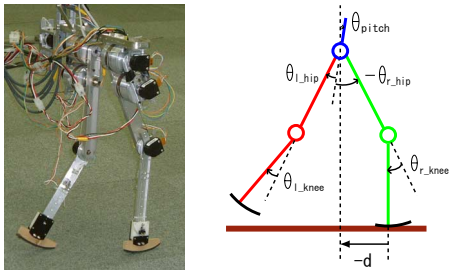


Fig. 1. Five link biped robot. Input state $\mathbf{x} = (d, \dot{d})$

First, we use a simulated 5 link biped robot (Fig. 1) to evaluate our proposed method. Physical parameters of the 5 link simulated robot in *TABLE I* are selected to model the actual biped robot fixed to a boom that keeps the robot in the sagittal plane (Fig. 1). Our biped has a short torso and round feet without ankle joints. For these bipeds, controlling biped walking trajectories with the popular ZMP approach [7], [22] is difficult or impossible, and thus an alternative method for controller design must be used.

In section II, we introduce our reinforcement learning method for biped walking. In section III, we show simulation results. In section IV, we present an implementation of the proposed method on the real robot, and then demonstrate that the robot acquires a successful walking pattern within 100 trials.

II. Poincaré-map-based reinforcement learning for biped locomotion

We improve biped walking controllers based on an approximated Poincaré map using a model-based reinforce-

TABLE I

PHYSICAL PARAMETERS OF THE FIVE LINK ROBOT MODEL

	trunk	thigh	shin
mass [kg]	2.0	0.64	0.15
length [m]	0.01	0.2	0.2
inertia ($\times 10^{-q}$ [$kg \cdot m^2$])	1.0	6.9	1.4

ment learning framework [3], [16]. The Poincare´ map represents the locus of intersection of the biped trajectory with a hyperplane subspace of the full trajectory state space. In our case, we are interested in the system state at two symmetric phase angles of the walking gait. Modulating via-points affects the locus of intersection and our learned model reflects this effect. Given a learned mapping, we proceed to learn a corresponding value function for states at phases $\varphi = \frac{1}{2}\pi$ and $\varphi = \frac{3}{2}\pi$ (Fig. 2), where we define phase $\varphi = 0$ as the left foot touchdown.

The input state is defined as $\mathbf{x} = (d, \dot{d})$, where d denotes the horizontal distance between the stance foot position and the body position (Fig. 1). We use the hip position as the body position because the center of mass is nearly coincident with the hips (Fig. 1). We use a human walking pattern in [5] as the nominal trajectory (Fig. 3). The action of the robot $\mathbf{u} = \boldsymbol{\theta}^{act}(\mathbf{x}) = (\theta^{act}_{hip}, \theta^{act}_{knee})$ modulates the via-points of the nominal trajectory at each joint:

$$\theta^i_{hip_vp} = \bar{\theta}^i_{hip_vp} + \theta^{act}_{hip}(\mathbf{x}) \quad (i = 1, \cdots, n^v_{hip}), \tag{1}$$

$$\theta^i_{knee_vp} = \bar{\theta}^i_{knee_vp} + \theta^{act}_{knee}(\mathbf{x}) \quad (i = 1, \cdots, n^v_{knee}), \tag{2}$$

where $n^v_{hip} = 1$ and $n^v_{knee} = 2$ denote the number of selected via-points, and $\bar{\theta}^i_{hip_vp}$ and $\bar{\theta}^i_{knee_vp}$ denote the nominal value of the selected via-points. Each selected via-point on a same joint is equally modulated by the control output θ^{act}.

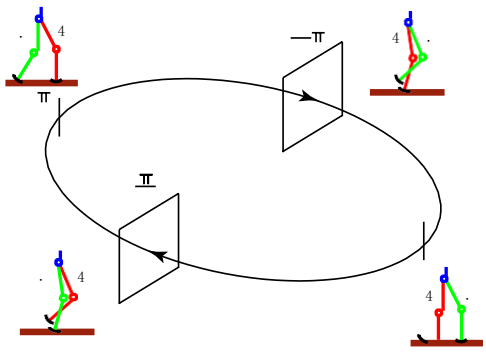


Fig. 2. Biped walking cycle: we update parameters and select actions at Poincare´ sections at phase $\varphi = \frac{\pi}{2}$ and $\varphi = \frac{3\pi}{2}$. L:left leg, R:right leg

A. Function approximator

We use Receptive Field Weighted Regression(RFWR) [15] as the function approximator for the policy, the value function and the estimated Poincare´ map. We approximate a target function $g(\mathbf{x})$ with

$$\hat{g}(\mathbf{x}) = \frac{\sum_{k=1}^{N_b} a_k(\mathbf{x}) h_k(\mathbf{x})}{\sum_{k=1}^{N_b} a_k(\mathbf{x})}, \tag{3}$$

$$h_k(\mathbf{x}) = \mathbf{w}_k^T \tilde{\mathbf{x}}_k, \tag{4}$$

$$a_k(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{c}_k)^T \mathbf{D}_k (\mathbf{x} - \mathbf{c}_k)\right), \tag{5}$$

where $\mathbf{c}_k$ is the center of the k-th basis function, $\mathbf{D}_k$ is the distance metric of the k-th basis function, N_b is the number of basis functions, and $\tilde{\mathbf{x}}_k = ((\mathbf{x} - \mathbf{c}_k)^T, 1)^T$ is the augmented state. The update rule for the parameter $\mathbf{w}$ is given by:

$$\Delta \mathbf{w}_k = a_k \mathbf{P}_k \tilde{\mathbf{x}}_k (g(\mathbf{x}) - h_k(\mathbf{x})), \tag{6}$$

where

$$\mathbf{P}_k \leftarrow \frac{1}{\lambda}\left(\mathbf{P}_k - \frac{\mathbf{P}_k \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T \mathbf{P}_k}{\frac{\lambda}{a_k} + \tilde{\mathbf{x}}_k^T \mathbf{P}_k \tilde{\mathbf{x}}_k}\right), \tag{7}$$

and $\lambda = 0.999$ is the forgetting factor.

In this study, we allocate a new basis function if the activation of all existing units is smaller than a threshold a_{min}, i.e.,

$$\max_k a_k(\mathbf{x}) < a_{\min}, \tag{8}$$

where $a_{\min} = \exp(-\frac{1}{2})$. We initially align basis functions $a_k(\mathbf{x})$ at even intervals in each dimension of input space $\mathbf{x} = (d, \dot{d})$ (Fig. 1) $[-0.2(m) \le d \le 0.2(m)$ and $-1.0(m/s) \le \dot{d} \le 1.0(m/s)]$. Initial numbers of basis

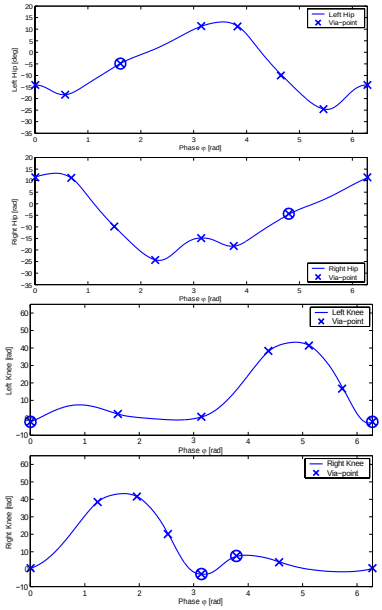


Fig. 3. Nominal joint-angle trajectories observed from a human walking pattern and detected via-points represented by cross (×). Manually selected via-points represented by circle (∘) are modulated by control output θ_{act}. Note that amplitude of the human walking pattern is multiplied by 0.7 to match the small size robot (Fig 1).

TABLE I

PHYSICAL PARAMETERS OF THE FIVE LINK ROBOT MODEL

	trunk	thigh	shin
mass $[kg]$	2.0	0.64	0.15
length $[m]$	0.01	0.2	0.2
inertia $(\times 10^{-4}\ [kg \cdot m^2])$	1.0	6.9	1.4

ment learning framework [3], [16]. The Poincaré map represents the locus of intersection of the biped trajectory with a hyperplane subspace of the full trajectory state space. In our case, we are interested in the system state at two symmetric phase angles of the walking gait. Modulating via-points affects the locus of intersection and our learned model reflects this effect. Given a learned mapping, we proceed to learn a corresponding value function for states at phases $\phi = \frac{1}{2}\pi$ and $\phi = \frac{3}{2}\pi$ (Fig. 2), where we define phase $\phi = 0$ as the left foot touchdown.

The input state is defined as $\mathbf{x} = (d, \dot{d})$, where d denotes the horizontal distance between the stance foot position and the body position (Fig. 1). We use the hip position as the body position because the center of mass is nearly coincident with the hips (Fig. 1). We use a human walking pattern in [5] as the nominal trajectory (Fig. 3). The action of the robot $\mathbf{u} = \boldsymbol{\theta}^{act}(\mathbf{x}) = (\theta^{act}_{hip}, \theta^{act}_{knee})$ modulates the via-points of the nominal trajectory at each joint:

$$\theta^i_{hip_vp} = \bar{\theta}^i_{hip_vp} + \theta^{act}_{hip}(\mathbf{x}) \quad (i = 1, \cdots, n^v_{hip}), \tag{1}$$

$$\theta^i_{knee_vp} = \bar{\theta}^i_{knee_vp} + \theta^{act}_{knee}(\mathbf{x}) \quad (i = 1, \cdots, n^v_{knee}), \tag{2}$$

where $n^v_{hip} = 1$ and $n^v_{knee} = 2$ denote the number of selected via-points, and $\bar{\theta}^i_{hip_vp}$ and $\bar{\theta}^i_{knee_vp}$ denote the nominal value of the selected via-points. Each selected via-point on a same joint is equally modulated by the control output θ^{act}.

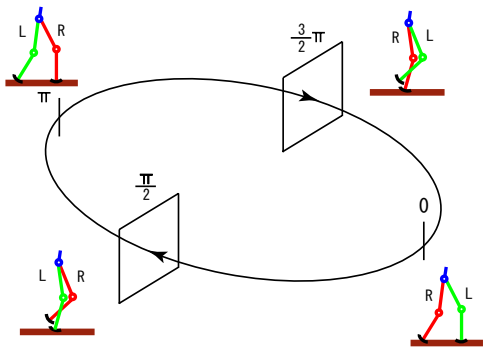


Fig. 2. Biped walking cycle: we update parameters and select actions at Poincaré sections at phase $\phi = \frac{\pi}{2}$ and $\phi = \frac{3\pi}{2}$. L:left leg, R:right leg

A. Function approximator

We use Receptive Field Weighted Regression(RFWR) [15] as the function approximator for the policy, the value function and the estimated Poincaré

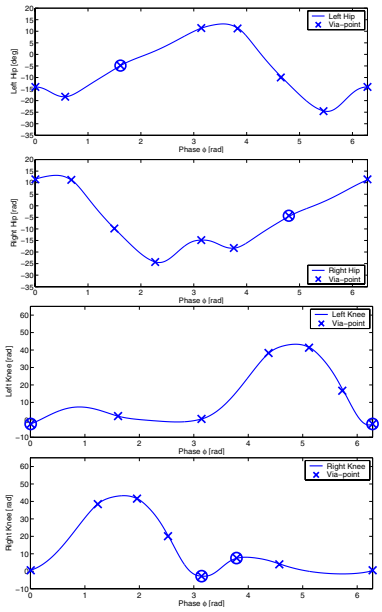


Fig. 3. Nominal joint-angle trajectories observed from a human walking pattern and detected via-points represented by cross (×). Manually selected via-points represented by circle (○) are modulated by control output θ_{act}. Note that amplitude of the human walking pattern is multiplied by 0.7 to match the small size robot (Fig 1).

map. We approximate a target function $g(\mathbf{x})$ with

$$\hat{g}(\mathbf{x}) = \frac{\sum_{k=1}^{N_b} a_k(\mathbf{x}) h_k(\mathbf{x})}{\sum_{k=1}^{N_b} a_k(\mathbf{x})}, \tag{3}$$

$$h_k(\mathbf{x}) = \mathbf{w}_k^T \tilde{\mathbf{x}}_k, \tag{4}$$

$$a_k(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{c}_k)^T \mathbf{D}_k (\mathbf{x} - \mathbf{c}_k)\right), \tag{5}$$

where $\mathbf{c}_k$ is the center of the k-th basis function, $\mathbf{D}_k$ is the distance metric of the k-th basis function, N_b is the number of basis functions, and $\tilde{\mathbf{x}}_k = ((\mathbf{x} - \mathbf{c}_k)^T, 1)^T$ is the augmented state. The update rule for the parameter $\mathbf{w}$ is given by:

$$\Delta \mathbf{w}_k = a_k \mathbf{P}_k \tilde{\mathbf{x}}_k (g(\mathbf{x}) - h_k(\mathbf{x})), \tag{6}$$

where

$$\mathbf{P}_k \leftarrow \frac{1}{\lambda}\left(\mathbf{P}_k - \frac{\mathbf{P}_k \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T \mathbf{P}_k}{\frac{\lambda}{a_k} + \tilde{\mathbf{x}}_k^T \mathbf{P}_k \tilde{\mathbf{x}}_k}\right), \tag{7}$$

and $\lambda = 0.999$ is the forgetting factor.

In this study, we allocate a new basis function if the activation of all existing units is smaller than a threshold a_{min}, i.e.,

$$\max_k a_k(\mathbf{x}) < a_{\min}, \tag{8}$$

where $a_{\min} = \exp(-\frac{1}{2})$. We initially align basis functions $a_k(\mathbf{x})$ at even intervals in each dimension of input space $\mathbf{x} = (d, \dot{d})$ (Fig. 1) $[-0.2(m) \leq d \leq 0.2(m)$ and $-1.0(m/s) \leq \dot{d} \leq 1.0(m/s)]$. Initial numbers of basis

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/225323133303011122