論文 / 著書情報
Article / Book Information

論題	モデルベース強化学習を用いた2足歩行運動の獲得
Title	
著者	森本淳, 中西淳, 遠藤玄, チェンゴードン, アトケソン クリストファー
Author	Jun Morimoto, Jun Nakanishi, Gen Endo, Gordon Cheng, Christpher Atkeson
掲載誌/書名	日本機械学会ロボティクス・メカトロニクス講演会'04講演論文集, Vol. , No. , pp. 1A1-L1-55
Journal/Book name	, Vol. , No. , pp. 1A1-L1-55
発行日 / Issue date	2004,
URL	http://www.jsme.or.jp/publish/transact/index.html
権利情報 / Copyright	本著作物の著作権は日本機械学会に帰属します。
Note	このファイルは著者（最終）版です。 This file is author (final) version.

モデルベース強化学習を用いた２足歩行運動の獲得

A Model-based Reinforcement Learning Algorithm For Biped Walking

○森本 淳（ATR）　　中西 淳（ATR，ICORP/JST）
遠藤 玄（SONY, ATR）　　Gordon Cheng （ATR）
Christopher G. Atkeson (CMU, ATR)

Jun MORIMOTO	ATR CNS, Kyoto Japan
Jun NAKANISHI	ATR CNS, ICORP JST, Kyoto Japan
Gen ENDO	Sony Corporation, Tokyo, Japan, ATR CNS, Kyoto, Japan
Gordon CHENG	ATR CNS, Kyoto Japan
Christopher G. ATKESON	CMU, Pittsburgh, USA, ATR CNS, Kyoto Japan

We propose a model-based reinforcement learning algorithm for biped walking in which the robot learns to appropriately place the swing leg. This decision is based on a learned model of the Poincare map of the periodic walking pattern. The model maps from a state at the middle of a step and foot placement to a state at next middle of a step. We also modify the desired walking cycle frequency and desired phase difference based on online measurements. We present simulation results, and are currently implementing this approach on an actual biped robot.

Key Words : Biped Walking, Reinforcement Learning, Walking period adaptation, Phase difference adaptation, Phase resetting

1. はじめに

本研究では，２足歩行運動の学習手法についての検討を行う．２足歩行運動の実現のためには，適切な歩行軌道の生成と，歩行タイミングの環境への適応が重要となる．ここでは，強化学習を用いた遊脚接地位置の獲得手法と，位相振動子を用いた遊脚接地タイミングの適応手法[4, 5]の提案を行う．２足歩行運動の獲得に強化学習を適用した例として[1, 2, 3]などがあるが，本研究では，学習によって獲得される軌道切断面の写像と価値関数を用いて，効率的な歩行運動学習を実現する．さらに，歩行周期および位相差を適応させることで，環境の変化に対応可能となることを示す．

提案手法を評価するため，図１に示した３リンクおよび５リンクのロボットモデルを用いる．それぞれのモデルは共に全長 0.4m，重さ 3.6kg である．５リンクモデルの物理パラメータは図１に示した実機をモデル化している．また，この２足歩行ロボットは胴体部が短く，足先が点接触および円弧接触であるため，２足歩行ロボットに広く用いられている，ZMP を基にした軌道計画および制御手法を適用することが困難である．本研究では，図１のような歩行ロボットモデルでも，歩行運動を獲得できることを示す．

2. 遊脚接地タイミングの適応

安定な２足歩行を実現するためには，適切なタイミングで遊脚の接地が行われなければならない．たとえば，適切な歩行タイミングは坂を降りる場合などは変化する．そこで本研究では，歩行ロボットの歩行周期と制御出力との位相差を環境に適応させる手法を提案する．

Fig. 1　Biped robot model

2·1 歩行周期および位相差の適応手法

ここでは，一歩の歩行に要した時間 T を用いて，次に示す目標歩行角周波数 ω^* を得る．

$$\omega^* = \frac{\pi}{T} \tag{1}$$

この目標値を用いて，遊脚接地時に次のように目標軌道の角周波数を更新する．

$$\hat{\omega}_{n+1} = \hat{\omega}_n + K_\omega(\omega^* - \hat{\omega}_n), \tag{2}$$

ただし，K_ω は角周波数適応ゲイン．ω_n は n ステップ後の目標軌道角周波数．この手法は著者らの研究[5]において，その有効性が確認されている．歩行周期の適応と同様に，位相をリセットすることが２足歩行運動において有効であることが実験的に示されている[4, 5]．従来手法では，制御器の出力とロボットの位相差がゼロとなるように位相リセットを行っている．このような手法は，ロボットが高精度で目標軌道を追従できることを仮定しているが，制御器の出力がトルクによって与えられる場合や，サーボのゲインが小さい場合は，出力軌道とロボットの歩行軌道に適切な位相差を与えることが必要となる．ここでは，適切な位相差への適応手法を提案する．ϕ^* を遊脚接地時の位相とし，次式を用いて位相リセットの目標値 $\bar{\phi}$ を更新する．その後，目標軌道の位相 ϕ を $\bar{\phi}$ でリセットする．

$$\begin{aligned} \bar{\phi}_{n+1} &= \bar{\phi}_n + K_\phi(\phi^* - \bar{\phi}_n) \\ \phi &\leftarrow \bar{\phi}_{n+1}, \end{aligned} \tag{3}$$

ここで，$\bar{\phi}_n$ は n ステップ後の目標位相差である．また，K_ϕ は位相差適応ゲインである．

3. 遊脚接地位置の学習

２足歩行運動の実現のためには，適切なタイミングで遊脚が接地するだけでなく，適切な位置に遊脚を接地さ

せる必要がある．ここでは，適切な遊脚接地位置を学習することで，安定な歩行を実現するための制御器の獲得を行う．また，歩行軌道は位相 ϕ 上の軌道として表される．

3·1 モデルベース強化学習

ここでは，モデルベース強化学習の枠組みを用いる[6, 7]．強化学習に用いるモデルとして，軌道切断面の写像を用いることにする．また，この写像は学習を通じて獲得される．価値関数の評価は位相 $\phi = \frac{\pi}{2}$，$\phi = \frac{3\pi}{2}$ において行われる．ただし図 2 に示すように，目標軌道の右脚接地時を位相 $\phi = 0$ とした．

3·1·1 軌道切断面写像の学習

軌道切断面における現在のロボットの状態と，遊脚接地位置を決定する膝目標関節角 θ_{act} から，半周期先の軌道切断面でのロボットの状態を予測する．たとえば，位相 $\phi = \frac{\pi}{2}$ における状態から位相 $\phi = \frac{3\pi}{2}$ における状態を予測する．遊脚接地時 $(\phi = 0, \pi)$ においては，床面との衝突により大きくロボットの状態が変化するため，ここでは位相 $\phi = \frac{\pi}{2}$，$\phi = \frac{3\pi}{2}$ における軌道切断面での状態を用いた．軌道切断面写像の学習は，パラメータベクトル $\mathbf{w}^m$ を持つ関数近似器によって実現される．

$$\hat{\mathbf{x}}_{\frac{3\pi}{2}} = \hat{\mathbf{f}}(\mathbf{x}_{\frac{\pi}{2}}, u_{\frac{\pi}{2}}; \mathbf{w}^m), \tag{4}$$

ただし，$\mathbf{x} = (d, \dot{d})$ はロボットの状態変数を表す．図 3 に示すように，d は支持脚の足先位置からロボットの腰位置の距離を表し，$u = \theta_{act}$ は遊脚膝関節の目標角を表す．

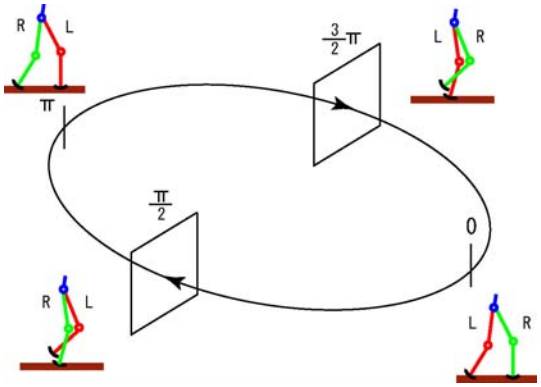

Fig. 2 Biped walking trajectory using four via-points: we update parameters and select actions at Poincare sections on phase $\phi = \frac{\pi}{2}$ and $\phi = \frac{3\pi}{2}$. L:left leg, R:right leg

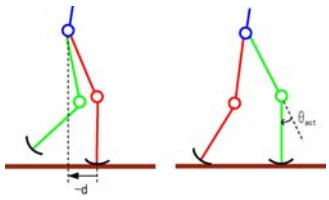

Fig. 3 (Left) Input state, (Right) Output of the controller

3·1·2 歩行軌道の設計

一周期の歩行軌道は，各関節に対して 4 つの経由点によって表現される（表 1 参照）．表 1 に示すように，$u = \theta_{act}$ は経由点として用いられる．ここでは，それぞれの経由点を内挿する手法として，関節角躍度最小軌道を用いた[8, 9]．生成される目標軌道を追従するために，以下に示す PD サーボを用いた．

$$\tau_j = k(\theta_j^d(\phi) - \theta_j) - b\dot{\theta}_j, \tag{5}$$

ただし，$\theta_j^d(\phi)$ は位相 ϕ における j 番目の関節角 θ_j（$j = 1, \cdots, 4$，図 1 参照）の目標軌道である．また，k，b はサーボゲイン，τ_j は j 番目の関節角でのトルク出力を表す．

Table 1 Target postures at each phase ϕ : θ_{act} is provided by the output of current policy. The unit for numbers in this table is degrees

	right hip	right knee	left hip	left knee
$\phi = 0$	-10.0	θ_{act}	10.0	0.0
$\phi = 0.5\pi$		θ_{act}		60.0
$\phi = 0.7\pi$	10.0		-10.0	
$\phi = \pi$	10.0	0.0	-10.0	θ_{act}
$\phi = 1.5\pi$		60.0		θ_{act}
$\phi = 1.7\pi$	-10.0		10.0	

3·1·3 報酬関数

ロボットは継続して歩き続けることができれば，正の報酬，転倒すれば負の報酬を得る．位相 $\phi = \frac{1}{2}\pi$(または $\phi = \frac{3}{2}\pi$) から位相 $\phi = \frac{3}{2}\pi$(または $\phi = \frac{1}{2}\pi$) への遷移の度にロボットは 0.1 の報酬を得る．ロボットが転倒すると-1.0 の負の報酬を得て，試行が与えられた初期位置から再開される．

3·1·4 評価関数の学習

強化学習の枠組みにおいては，累積報酬の期待値を最大化するような制御器の学習を行う．ここで，制御器 μ の価値関数を次のように定義する．

$$V^{\mu}(\mathbf{x}(t)) = E[r(t+1) + \gamma r(t+2) + \gamma^2 r(t+3) + ...], \tag{6}$$

ただし，$r(t)$ は時刻 t における報酬であり，γ $(0 \leq \gamma \leq 1)$ は割引率を示す．本研究で提案する枠組みでは，位相 $\phi = \frac{1}{2}\pi$ および $\phi = \frac{3}{2}\pi$ においてのみ価値関数の更新を行う．よって提案手法は，セミマルコフ決定過程での学習手法となる．価値関数は次のようにパラメータベクトル $\mathbf{w}^v$ を持つ関数近似器によって近似される．

$$\hat{V}(t) = \hat{V}(\mathbf{x}(t); \mathbf{w}^v). \tag{7}$$

式 (6) からの誤差を考慮することで，次に示す TD 誤差が定義できる．

$$\delta(t_T) = \sum_{k=t_T+1}^{t_{T+1}} \gamma^{k-t_T-1} r(k) + \gamma^{t_{T+1}-t_T} \hat{V}(t_{T+1}) - \hat{V}(t_T), \tag{8}$$

ただし，t_T は位相が $\phi(t_T) = \frac{1}{2}\pi$ または $\phi(t_T) = \frac{3}{2}\pi$ であるときの時間である．よって，価値関数の更新則は，

$$\hat{V}(\mathbf{x}(t_T); \mathbf{w}^v) \leftarrow \hat{V}(\mathbf{x}(t_T); \mathbf{w}^v) + \beta\delta(t_T), \qquad (9)$$

となる．ただし $\beta = 0.2$ は学習率である．

3·1·5　歩行制御器の学習

ここでは次に示す確率的な制御器を用いる．

$$\mu(\mathbf{u}(t_T)|\mathbf{x}(t_T)) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\mathbf{u}(t_T) - \mathbf{A}(\mathbf{x}(t_T); \mathbf{w}^a))^2}{2\sigma^2}\right), \qquad (10)$$

ただし，$\mathbf{A}(\mathbf{x}(t_T); \mathbf{w}^a)$ は制御器の平均出力を示し，パラメータベクトル $\mathbf{w}^a$ を持つ関数近似器によって表される．よって制御出力の実現値は次のようになる．

$$\mathbf{u}(t_T) = \mathbf{A}(\mathbf{x}(t_T); \mathbf{w}^a) + \sigma\mathbf{n}, \qquad (11)$$

ただし，$\mathbf{n} \sim N(0, 1)$ であり，$N(0, 1)$ は平均 0，分散 1 の正規分布を表す．制御器の更新則は，価値関数と軌道切断面写像を用いることで，次のように与えられる．

1. 状態 $\mathbf{x}(t_T)$ における価値関数の勾配 $\frac{\partial V}{\partial \mathbf{x}}$ を求める．
2. 半周期前の切断面での状態 $\mathbf{x}(t_{T-1})$ と平均出力 $\mathbf{u} = \mathbf{A}(\mathbf{x}(t_{T-1}); \mathbf{w}^a)$ における軌道切断面写像の勾配 $\frac{\partial \mathbf{f}}{\partial \mathbf{u}}$ を求める．
3. 次のように制御器を更新する．

$$\mathbf{A}(\mathbf{x}; \mathbf{w}^a) \leftarrow \mathbf{A}(\mathbf{x}; \mathbf{w}^a) + \alpha \frac{\partial V(\mathbf{x})}{\partial \mathbf{x}} \frac{\partial \mathbf{f}(\mathbf{x}, \mathbf{u})}{\partial \mathbf{u}}, \qquad (12)$$

ただし，$\alpha = 0.2$ は学習率である．

4.　シミュレーション結果

4·1　3リンクロボット モデルを用いた歩行周期・位相差の適応

ここでは，図 1 に示した 3 リンクのロボットモデルを用い，歩行周期および位相差の適応実験を行う．サイン関数によって目標軌道を生成し，PD サーボを用いて目標軌道に対する追従を行った．

$$\tau_l = k(a\sin\phi - \theta_l) - b\dot{\theta}_l \qquad (13)$$

$$\tau_r = k(-a\sin\phi - \theta_r) - b\dot{\theta}_r, \qquad (14)$$

ただし，τ_l，τ_r は左右腰関節へのトルク入力．$k = 5.0$，$b = 0.1$ はサーボゲイン，θ_l，θ_r は左右腰関節角を表す．位相 ϕ は，$\phi = \hat{\omega}_n t$ によって与えられる．初期周期を $T = 0.63sec$(初期周波数 $\omega_0 = 10rad/sec$) とした．また，初期目標位相差を $\bar{\phi} = 1.0rad$ とした．周期適応ゲインを $K_\omega = 0.3$ とし，位相差適応ゲインを $K_\phi = 0.3$ とした．初期状態として，水平進行方向に $0.2m/s$ の初速度を与えた．まずはじめに，適応手法を用いない場合のシミュレーション実験を行った．図 4 に示すように 3 リンクロボットモデルは，1.0° の坂を歩行することができたが，4.0° の坂は安定して下ることができなかった．しかし，式 (2)，(3) に示した適応手法を用いることで，図 5 に示すように，1.0° と 4.0° の両方の坂を安定して歩行することができた．これは，適応手法によって，適切な歩行周期と位相差が得られたためと考えられる．図 6 上段に目標軌道の歩行角周波数を，図 6 下段に目標位相差をそれぞれの坂に適応させた場合について示した．

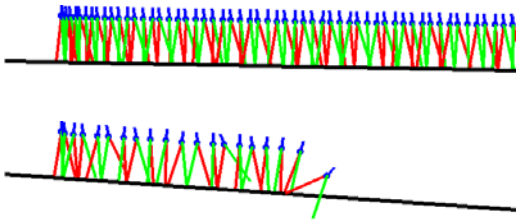

Fig. 4　Biped walking pattern without timing adaptation: (Top) 1.0° downward slope, (Bottom) 4.0° downward slope

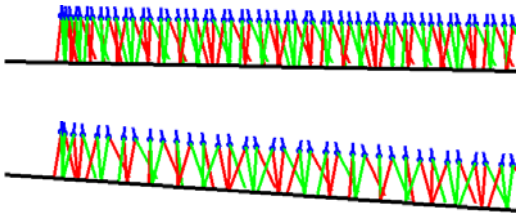

Fig. 5　Biped walking pattern with timing adaptation: (Top) 1.0° downward slope, (Bottom) 4.0° downward slope

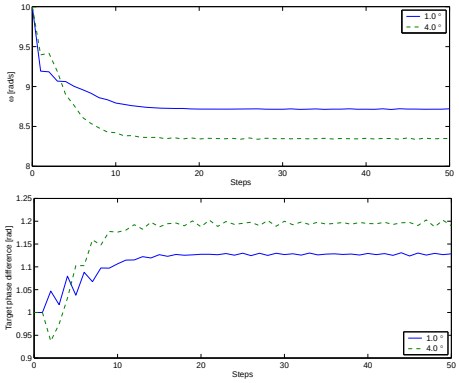

Fig. 6　(Top)Walking frequency of the target trajectories. (Bottom)Desired phase difference $\bar{\phi}$.

4·2　モデルベース強化学習を用いた遊脚接地位置の獲得

ここでは，提案したモデルベース強化学習法を図 1 に示した 5 リンクのロボットモデルに適用した．

試行は 50 歩の歩行に成功するか，または転倒した場合に打ち切った．歩行周期を $T = 0.79sec$ ($\omega = 8.0[rad/sec]$) に設定した．図 7 上段に学習前の歩行軌道を，図 7 中段に 30 試行後の歩行軌道を示した．図 8 左図はそれぞれの試行での累積報酬を示している．ここでは，はじめて 50 歩の歩行に成功した時に，歩行獲得と定義する．5 回の実験を行った結果，平均 80 試行で歩行運動を獲得した．図 7 下段に獲得された歩行軌道を示す．また，獲得された価値関数を図 8 右図に示した．

4·3　歩行周期・位相差適応手法の獲得された制御器への適用

ここでは，前節で獲得された制御器に対して，提案した歩行周期・位相差適応手法を適用する．初期目標位相差を $\bar{\phi} = 1.0rad$ とした．角周波数適応ゲインは $K_\omega = 0.3$,

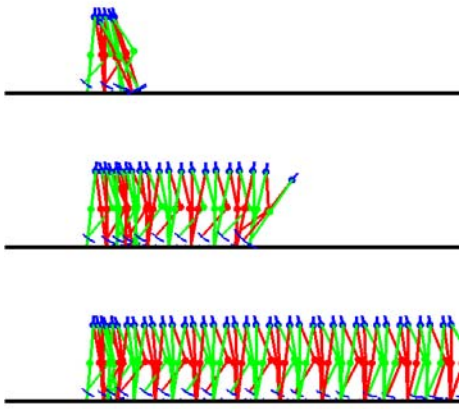

Fig. 7 Acquired biped walking pattern: (Top)Before learning, (Middle)After 30 trials, (Bottom)After learning

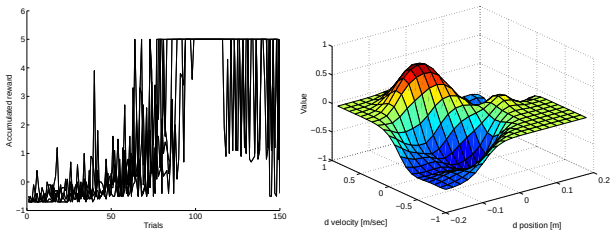

Fig. 8 (Left)Accumulated reward at each trial: Results of five experiments. (Right)Shape of acquired value function

位相適応ゲインは $K_\phi = 0.3$ とした．

ここでは，1.0° の下り坂に獲得された制御器を適用した．図 9 上段に，適応手法を用いない場合の結果を示した．数歩で転倒していることがわかる．一方，図 9 下段に示すように，適応手法を用いることで安定な歩行運動が実現ができていることがわかる．

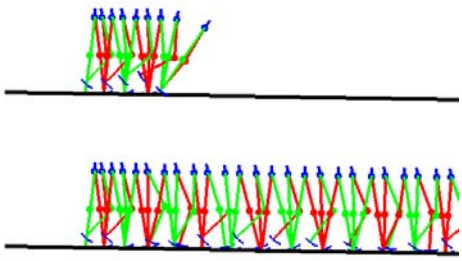

Fig. 9 Biped walking pattern with timing adaptation on downward slope: (Top)Without timing adaptation, (Bottom)With timing adaptation

4·4 獲得された制御器の安定性

軌道切断面写像を用いて，獲得された制御器の安定性の解析を行う．軌道切断面における安定歩行軌道付近での線形化された写像を表すヤコビ行列 J の固有値が $|\lambda_i(j)| < 1 (i = 1, 2)$ であるかどうかで，安定性を評価する．3·1 節で提案したモデルベース強化学習では，状態に関して連続の関数近似器を用いているため，次のようにヤコビ行列 J を求めることができる．

$$J = \frac{d\mathbf{f}}{d\mathbf{x}} = \frac{\partial \mathbf{f}}{\partial \mathbf{x}} + \frac{\partial \mathbf{f}}{\partial \mathbf{u}} \frac{\partial \mathbf{u}(\mathbf{x})}{\partial \mathbf{x}}. \tag{15}$$

図 10 は試行ごとの固有値の平均を示している．学習が進むにつれて固有値が減少し，最終的に $|\lambda_i(j)| < 1$ となっていることがわかる．

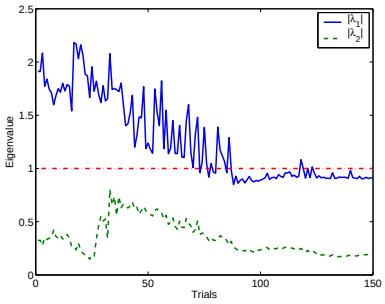

Fig. 10 Averaged eigenvalue of Jacobian matrix at each trial

4·5 まとめ

本研究では，適切な遊脚接地位置と遊脚接地タイミングを獲得するための手法を提案し，その有効性を検証した．今回は設計された歩行軌道を基に学習を行ったが，今後は人間の歩行計測データを基にした学習を行う予定である．さらに，今後提案手法の実機への適用を行う．

参考文献

(1) 中村泰, 佐藤雅昭, 石井信. 神経振動子ネットワークを用いたリズム運動に対する強化学習法. 電子情報通信学会論文誌, No. 3, pp. 893–902, 2004.

(2) H. Benbrahim and J. Franklin. Biped dynamic walking using reinforcement learning. *Robotics and Autonomous Systems*, Vol. 22, pp. 283–302, 1997.

(3) C. Chew and G. A. Pratt. Dynamic bipedal walking assisted by learning. *Robotica*, Vol. 20, pp. 477–491, 2002.

(4) K. Tsuchiya, S. Aoi, and K. Tsujita. Locomotion control of a biped locomotion robot using nonlinear oscillators. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1745–1750, Las Vegas, NV, USA, 2003.

(5) J. Nakanishi, J. Morimoto, G. Endo, G. Cheng, S. Schaal, and M. Kawato. Learning from demonstration and adaptation of biped locomotion. *Robotics and Autonomous Systems (to appear)*, 2004.

(6) K. Doya. Reinforcement Learning in Continuous Time and Space. *Neural Computation*, Vol. 12, No. 1, pp. 219–245, 2000.

(7) R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, 1998.

(8) T. Flash and N. Hogan. The coordination of arm movements: An experimentally confirmed mathematical model. *The Journal of Neuroscience*, Vol. 5, pp. 1688–1703, 1985.

(9) Y. Wada and M. Kawato. A theory for cursive handwriting based on the minimization principle. *Biological Cybernetics*, Vol. 73, pp. 3–15, 1995.