1D4-02

強化学習による多自由度２足歩行ロボットの制御

Controlling Multi-degree of Freedom Biped Robot by Reinforcement Learning

村田 栄理*1 Hidemasa Murata　浅井 孝宣*1 Takanobu Asai　佐久間 淳*1 Jun Sakuma　小林 重信*1 Shigenobu Kobayashi

*1東京工業大学 大学院総合理工学研究科
Interdisciplinary Graduate School of Science and Eng., Tokyo Institute of Technology

Recently, great efforts have been paid to the control problem of humanoid robots, the control of these robots is difficult because of the following properties; (1)multi-degree of freedom (2)complex structure (3)instability. However, the motion obtained by these methods is inferior to the human locomotion from the view point of natural locomotion. In our model of humanoid robot have fourteen degrees of freedom. In this paper, we present a new reinforcement learning based on a demonstration of human walking where appropriate amounts of the correction of each time step is output.

1. 背景と目的

現在, 制御が困難とされている多自由度ロボットが多く存在する. ヒト型ロボットがその例の一つとして挙げられる. このロボットは, 機体が床に固定されていない為, 極めて不安定であり, より綿密な制御器の設計が必要とされている. そして, この制御問題をベンチマークとして多くの研究が現在なされている. 代表的なアプローチとして ZMP(Zero Mometnt Point)[1] に基づくアプローチがある. このアプローチでは, ZMP を安定規範として ZMP の位置を支持脚多角形から飛び出さないように各関節を制御することで転倒を回避した運動を実現する. このアプローチを用いてロボットに理想の運動を行わさせる際は, 各関節の制御量を決定するときに, あらかじめロボットの機構を正確に把握した上で, 逆運動学問題を解き理想の制御量を算出する必要がある. しかし, 逆運動学を用いた際, 特異点を回避するため膝を曲げたままの歩行パターンになってしまうこと, エネルギー消費量が大きいこと [2] などが指摘されており, その歩容はヒトが行う自然な動作とは異なっている.

一方, ヒトらしい歩行を目指したものとして, Nakanishi ら [3] は, ヒトの関節運動を位相振動子としてエンコードし, この振動子とロボットの身体ダイナミクスとの引きこみ現象を, 足の接地情報のフィードバックによって実現することで, ヒト歩行によく似た自然な歩行生成に成功している. しかし, この手法ではロボットを振子系として扱いやすい単純な 5 リンクモデルとして定式化しており, 足関節を持たず足裏が円弧で接地するという構造制約を持つため, 構造においてヒトらしさを欠くという問題点がある. さらに, ヒトの下半身の自由度はもっと多く, 少ない自由度で設計されたモデルはヒトに比べて運動性能でおとる.

そこで, 足首関節を備えたヒトらしいモデルにおいて, ヒトのように自然な歩行を実現することを目的とした浅井ら [4] の研究に着目する. 浅井らはヒトらしい歩行を生成するために, ヒトに近い身体構造を持った 6 自由度ロボットモデルに対し, 強化学習を用いてヒトの歩行軌道例に修正を加えることで矢状面における安定歩行を実現した. 本研究では, 下半身の自由度を 14 自由度とすることで, 矢状面制約を取り除いた 3 次元環境下において, よりヒトに近い運動性能の実現を試みる.

以下, 本稿は次のように構成される. 2 章モデル設定では, 浅井モデルに基づく制御対象の身体機構と, 出力に用いる関節軌道 (基本関節軌道) の設定について述べる. 3 章強化学習を用いた制御では, 強化学習を用いて基本関節軌道に修正を加える方法について説明する. 4 章実験では, シミュレーションモデルを用いて実験を行い, 強化学習を用いたフィードバック機構により高い運動性能が得られたことを示す. 5 章おわりにでは, 研究成果をもとに, 今後の課題について述べる.

表 1: 関節の可動方向と範囲

股関節	左右	-10 度 ～ 10 度
股関節	前後	-45 度 ～ 135 度
膝関節	前後	-5 度 ～ 145 度
足首関節	前後	-75 度 ～ 30 度
足首関節	左右	-20 度 ～ 20 度
つま先関節	前後	0 度 ～ 60 度
股関節	回転	-30 度 ～ 30 度

2. モデル設定

本研究では, ヒトの基本特性に関するデータベース [5] を参照して浅井らによって構築された, 身長 173cm 程のロボットモデルを扱う (図 1). また, 運動を歩行に限っている為, 上半身の自由度は考慮しないものとし, モデルの全自由度は 14 自由度とした (表 1). さらに, モデルは高精度な 3 次元動力学シミュレータ VORTEX[8] を用いて構築する. 各関節には角度制御のサーボモータを備えている. このロボットに搭載されているサーボモータは, 10ms 毎に目標値を与え PID 制御を行うと想定している.

ロボットの両足裏には, 踵とボール (親指の付け根よりやや土踏まずより部分) の両方が接地したときに反応する接触センサが備えてあり, 両足それぞれの接地状態を知ることができる. また体幹には, 鉛直軸からの傾斜角度を前後・左右それぞれに対し測定する角度センサと, 鉛直軸まわりの角速度を測定するジャイロセンサが備えられ, 上半身の情報を知ることができる. 以上,5 個のセンサが搭載されている.

連絡先: 村田 栄理, 東京工業大学大学院 総合理工学研究科 知能システム科学専攻, hidemasa@fe.dis.titech.ac.jp

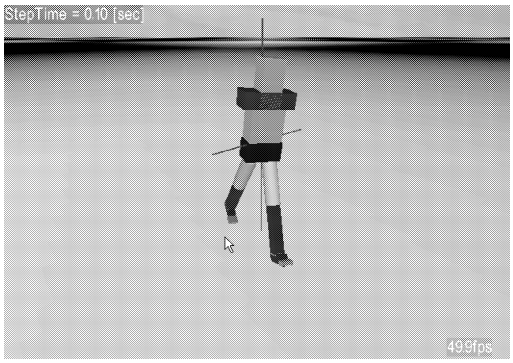

図 1: モデル

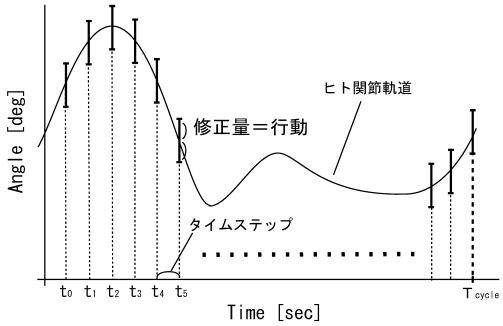

図 2: 膝関節の軌道データ（1 周期）

3. 強化学習を用いた制御

未知環境において, 試行錯誤を通じ求める運動性能の評価値を最大化する学習アルゴリズムのフレームワークに強化学習がある. 強化学習は, 獲得した政策への評価を明示的に与えることで政策を学習するのではなく, 状態毎の報酬を与え環境とのインタラクションを通じてより多くの報酬を獲得するような政策の学習を目指す. ここで, 政策とは学習アルゴリズム本体 (エージェント) が持つ状態観測と行動の入出力関係であり, 一般にロボットに適用する場合, 入力がセンサ情報, 出力が制御量の決定に影響を及ぼす情報となっている. その為, 評価関数の設定が困難な運動の終端状態を定義できない場合において有効である. また, 強化学習は, 制御対象の身体自体も環境の一部と考える為, ロボットの身体性も考慮に入れた最適な制御規則の獲得を期待できる. 強化学習の制御器への適用にあたって, 本研究ではセンサの値を基にタイムステップ毎に基本関節軌道に与える修正量の出力を学習する (図 2).

また身体機構をヒトに近づけたことで, ヒトの関節軌道データとロボットがヒトのように歩く際の軌道は近いと思われる. そこで, ロボットの基本関節軌道にはヒトの関節軌道データ [7] を用いる. 詳細としては, 歩行時に駆動する主要な関節として, 矢状面で駆動する股関節・膝関節・足首関節, 左右合わせて 6 自由度に対し, ヒトの関節データを用いる (図 3). それ以外の 8 自由度に対しては適当に設定した一定値を基本軌道として用いる.

3.1 強化学習の定式化

前節の議論に基づき、強化学習の定式化を行う. エージェントは状態入力 $s(t)$ として, 左右の足裏の接地センサに始まる各

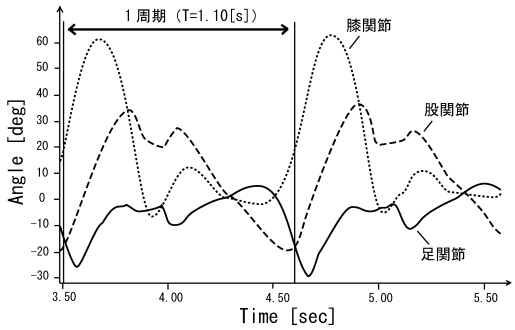

図 3: 主要 6 関節の基本関節軌道

表 2: Actor-Critic のパラメータ設定

	Actor	Critic
学習率	0.01	0.1
適正度の履歴の割引率	1.0	1.0
報酬の割引率	0.95	

5 センサ $s_1, s_2, \ldots, s_5$, 周期時刻インデックス j を観測する.

$$s(t) = (s_1, ..., s_5\, j)$$

周期時刻インデックス j は, 歩行開始時に $j = 0$ とし, 意思決定ステップ時間が進むにつれ $[0, T_{cycle}/t_{step} - 1]$ の範囲の整数を周期的にとる。ここで, T_{cycle} はヒト基本軌道の周期とする. また, 接地センサは $0, 1$ の離散値, 傾斜角度等その他のセンサは $[-\pi/2, \pi/2]$ の連続値である.

行動出力 $a(t)$ は, 各時刻でのヒト関節軌道に対する角度修正量ベクトルである.

$$a(t) = (\Delta a_0, \Delta a_1, \ldots, \Delta_5)$$

ただし, 角度修正量 Δa_i は, $[-\Delta_{range}, \Delta_{range}]$ の範囲の連続値とする. Δ_{range} は行動空間の大きさを決定するパラメータである.

エージェントの目的は転倒せずに継続歩行を行うことである. そこで報酬として, ロボットが転倒したときには負の転倒罰 $R_{falldown}$ を与える. なお, 体幹の傾斜角度がある閾値 θ_{thre} を超えた場合はそれ以上転倒回避できないと判断し, その時点で転倒と判断する。転倒罰が得られた場合はエピソードをリセットし, 学習を再スタートする.

報酬には, 転倒罰に加えて, 体幹の傾斜角度に関する罰も与える. これにより, 体幹の振れができるだけ少ない歩行の獲得を目指す. よって, 報酬 $r(t)$ は以下のように表される.

$$r(t) = \begin{cases} -|s_i| & if |s_i| < \theta_{thre} \\ R_{punishment} & if \quad |s_i| \geq \theta_{thre} \end{cases}$$

強化学習アルゴリズムとしては,“適正度の履歴に基づく Actor-Critic” [9] を用いる. 図 4 に, Actor-Critic の一般的な枠組みを示す. この手法は, 連続な状態行動空間を扱える, 非マルコフ性に対処できるなどの特徴をもつので, 本研究が対象とする学習問題に適していると考えられる.

また将来的に実装することを想定した場合, サーボモータに目標角度を与える度に, 学習により修正量を決定したのでは, 計算コストが必要以上にかかりすぎる恐れがある上, サーボモー

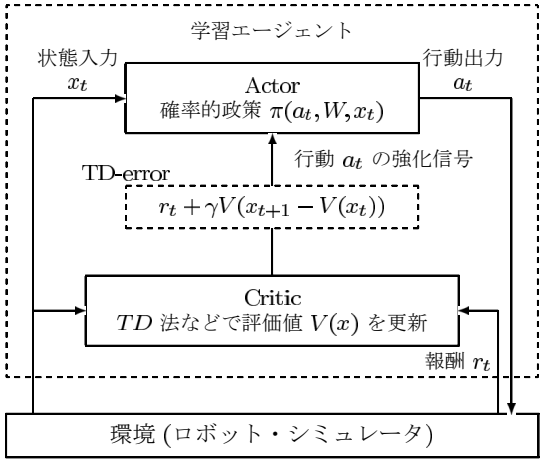

図 4: Actor-Critic の一般的枠組

タに指令値を送るタイミングまでに計算が間に合わない可能性もある. そこで, 制御量をロボットに送る意思決定のタイミングは 100ms とし, この値を次の意思決定として, 1 つ前の意思決定のとの差を 10 等分して, サーボモータに 10ms 毎のタイミングで指令値を送ることにした.

4. 実験

4.1 目的と設定

3 次元環境下における歩行中の転倒回避の行動を実現することを目的とする. 実験は, シミュレーション上で行い提案手法の有効性を検証する. 動力学シミュレータ上に構築したモデルに対し, 制御指令を送ってロボットを制御する, その結果得られるセンサ情報と報酬を学習エージェント側が観測し, 学習エージェントは内部パラメータを更新する. また, 学習エージェントの実装については, 文献 [10] の手法を参考にした. また, 先行研究である浅井モデルの実験設定に対し, 本実験は関節の数を 6 関節から 14 関節に増やし, その増やした関節の基本関節データは未知なので初期値のまま一定として実験した. 実験設定は, 角度修正量の範囲 $\Delta_{range} = 5$ 度, 転倒罰 $R_{falldown} = -5$, 転倒を判断する前後の傾斜角度の閾値 $\theta_{thre} = 0.209$ ラジアン (12 度) とした.

4.2 実験結果

実験を 23,000,000step 行った結果, 3 次元空間における歩行運動を獲得した (図 5,6). また, 図 7 のような学習曲線が得られた. この図の縦軸は 100step 毎の平均獲得報酬である. 実験では, 報酬を負の値 (罰) のみで与えたため 0 を超えることはないが, 図より学習曲線が 0 に近づいて収束していることがわかる. 結果, より罰を回避するように関節軌道に修正量を加えるように学習していることが確認された. ちなみに, 基本関節軌道にヒトのデータを用いなかった足首関節の左右駆動部の軌道 (図 8) を調べてみると, 周期的に修正量を加えられており, その他の関節においても同様であった. この結果からも, 周期状態に応じて適切な修正量が加えられていることがいえる.

また, 学習により得られた政策を用いた歩行において転倒回数を測定したところ, 500,000step 中 15 回の転倒があった. これは 1step あたりの罰の期待値が -1.5×10^{-4} となるため, 政策が得る罰はほとんど体幹の傾斜によるものであるとみなせる. 次に, 3 次元空間にロボットを配置してからの歩行時における, 体幹の前後の傾斜角を測定した (図 9). その結果, 歩行開始時に体幹は大きくぶれるが, その振れ幅を小さくするように歩行に移っていることが観察された. また, ヒト歩行における体幹の前後の振幅 2 度 (≈ 0.0348 ラジアン) であることから, 学習結果による安定歩行時の体幹の前後の振幅 0.05 ラジアンは全てを剛体のパーツで構成したロボットにしては妥当であると考えられる. よって, 収束した学習結果は極めて妥当だと言える.

得られたロボットの行動は, 当初求めていた 3 次元空間における転倒を回避した歩行を実現した. しかしながら得られた歩行は, 同じ箇所を何度も周回する歩行であり, 直進歩行時のヒト関節軌道を用いているにもかかわらず直進歩行が実現されていなう (図 10). 図 10 は, 右上から歩行を開始させ, 左下に到達した際にシミュレーションを止めたときの歩行の軌跡を上空から見たものである. また学習を継続させた場合, パラメータの探索を継続する為, 安定歩行中に転倒することもあった. その場合, 一度転倒すると安定歩行をなかなか実現することができなかった.

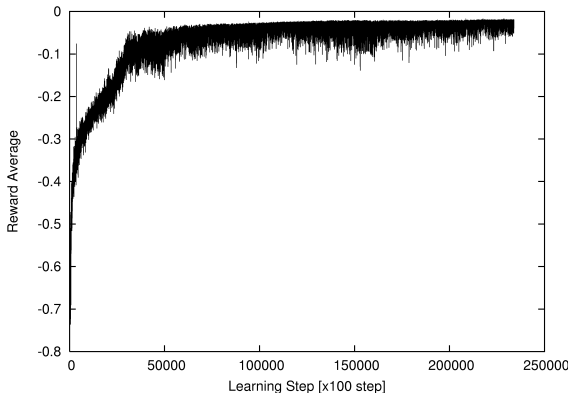

図 7: 学習曲線

4.3 考察

実験では, 直進歩行時のヒト関節軌道をベースに, 体幹の前後の傾きを運動評価の対象にすることで, 安定した歩行を実現する修正量を学習により与えることができた. しかし, 実際には周回しながらの歩行が得られ, 基本軌道を利用する際に想定していた行動とは大きく異なった. これは, 一度右にバランスを崩すと右方向に足を着き続けることが体幹の安定につながる為だと考えられる. そのため歩行の運動評価に, 体幹の安定化の他に直進性の評価を含む必要があると考えられる. さらに, ヒトは視覚情報を用いており目標方向へ移動することが歩行の最大の目的である為, ヒトらしい運動を行わせるためには, 目標方向への到達度をより明示的に与えるべきである.

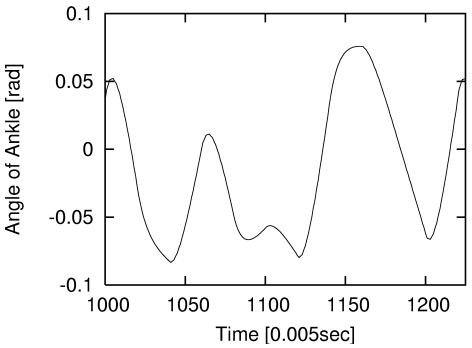

図 8: 学習後の足首関節角 (左右駆動部) の軌道

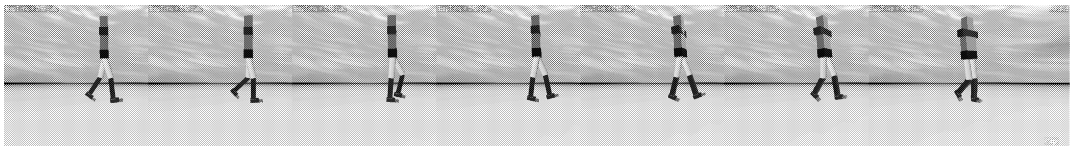

図 5: 学習により得られた歩行 (横から観測)

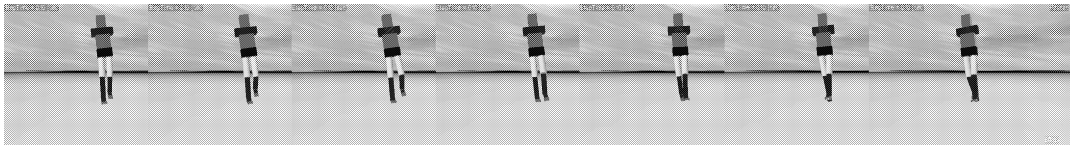

図 6: 学習により得られた歩行 (正面から観測)

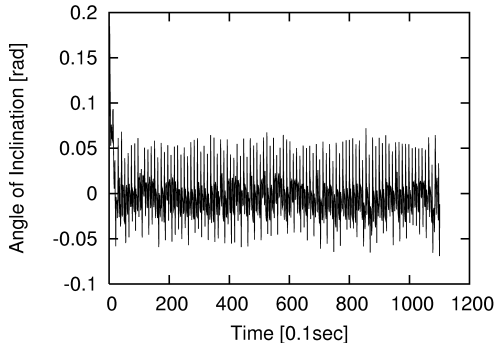

図 9: 体幹の前後の傾斜角

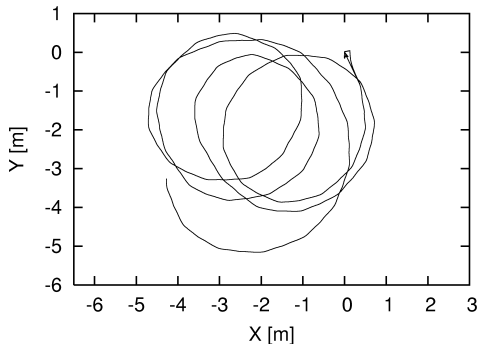

図 10: 学習後の歩行の軌跡 (上空から観測)

学習継続時に安定歩行中に転倒する原因は, さらなる最良の内部パラメータを学習器が探索するため, ランダムな行動をとることがあるためである. その際, 転倒後, 再び歩行をさせようとしたときになかなか安定歩行に移れない原因として, 安定歩行中の過学習が考えられる. 安定歩行中は不安定状態に遷移しない為, 不安定状態における転倒罰回避よりも, より体幹を安定化させることを追求した内部パラメータの更新が行われるため, 不安定状態時に学習したパラメータが失われ, 転倒後なかなか安定歩行をすることができなくなったと考えられる.

5. おわりに

実験では, 体幹の前後の傾きを運動評価の対象にしたが, 実際には目的の進行方向に対する直進性やエネルギー効率など, 本来のヒトらしい歩行を目指すためには更なる詳細な設定を考慮しなければならない. その際に, 考慮すべき問題として, 学習の多目的最適化問題や, 獲得政策のロバスト性などがある. また, 安定状態, 不安定状態等における制御器の使い分けの必要性も挙げられる. 本研究結果のように環境が変わる毎に制御対象が学習をしているのでは非効率であり, 今までの経験から得られた政策の効果的な利用が必要である.

今後の課題としては, 本研究ではヒト関節データを基本関節軌道として用い, データのない関節においてはゼロベースで学習を行ったが, 従来手法に比べてその妥当性については未だ検討の余地がある. そこで, 逆運動学を用いて制約を加えるなど今後は設計者が明示的に与えることができる情報は極力与え, それ以外の設定にのみ学習器を使う方針である.

参考文献

[1] M.Vukobratović, B.Borovac, D.Surla, D.Stokić: Biped Locomotion–Dynamics, Stability, Control and Application, Springer-Verlag (1990)

[2] S.Collins, M.Wisse, A.Ruina: A Three Dimentional Passive-Dynamic Walking Robot with Two Legs and Knees, *Proc. of the International Journal of Robotoics Research*, Vol.20, No.7, pp.607-615 (2001)

[3] J.Nakanishi, J.Morimoto, G.Endo, G.Cheng, S.Schaal, M.Kawato: Learning from Demonstration and Adaptation of Biped Locomotion with Dynamical Movement Primitives, *Workshop on Robot Programming by Demonstration, IEEE/RSJ International Conference on Intelligent Robots and Systems* (2003)

[4] 浅井 孝宣, 佐久間 淳, 小林 重信: ヒト関節軌道データを考慮した歩行ロボットの学習, 第 17 回自律分散システムシンポジウム (2005)

[5] 河内 まき子, 持丸 正明, 岩澤 洋, 三谷 誠二: 日本人人体寸法データベース 1997-98, 通商産業省工業技術院くらしと JIS センター (2000)

[6] 独立行政法人 製品評価技術基盤機構: 人間特性データベース, http://www.tech.nite.go.jp/human/

[7] 江原 義弘, 山本 澄子: ボディダイナミクス入門 歩き始めと歩行の分析, 医歯薬出版株式会社 (2002)

[8] http://www.cm-labs.com/

[9] 木村 元, 小林 重信: Actor に適正度の履歴を用いた Actor-Critic アルゴリズム, 人工知能学会誌, Vol.15, No.2, pp.267–275 (2000)

[10] 木村 元, 山下 透, 小林 重信: 強化学習による 4 足ロボットの歩行動作獲得, 電気学会 電子情報システム部門誌, Vol.122-C, No.3, pp.330–337 (2002)