様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年 6 月 4 日現在

機関番号：３２６８２
研究種目：基盤研究（C）
研究期間：2009～2011
課題番号：21560275
研究課題名（和文）腱駆動２足歩行ロボットの行動戦略・関節剛性の強化学習

研究課題名（英文）Reinforcement Learning of Action Strategies and Joint Stiffness of Tendon-driven Biped Robot

研究代表者
小林　博明（KOBAYASHI HIROAKI）
明治大学・理工学部・教授
研究者番号：60130811

研究成果の概要（和文）：
本研究では、様々な状況でどのように行動すればよいかを、罰と報酬を用いてロボット自身に学習させる手法を検討し、それをロボットサッカーゲームでの行動学習と２足歩行ロボットの歩行機能の学習に適用した。その際、実際のロボットに適用できるように、罰を与える基準の決定法、十分学習の進んだ状態は固定状態（一定の行動戦略を使用する状態）とするなど、学習の効率化を図った。また、人間と同様にモータ（筋肉）とワイヤー（腱）で駆動される２足歩行ロボットの機構と制御について研究した。２足歩行ロボットの腱には約 400N(40kgf)の力が加わるため、壊れにくい剛性調整装置を用い、腱張力の制御を行った。

研究成果の概要（英文）：
In this research, a learning method for robots to learn appropriate actions by profits and penalties given from the environment was developed and applied to action learning in the robotic succor game and walking movement of a biped robot. To apply it to the real robots and to improve the efficiency, a method to decide the criterion for penalties was considered and states in that the robot already had learned sufficiently were treated as a fixed-mode state (deterministic action strategy is used). Furthermore, the mechanism and control of a biped robot driven with motors (muscles) and wires (tendons) were considered. The tensile force control of tendons was done with robust stiffness-adjustable device, since tensile force up to 400N (40kgf) is expected during walking.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	500,000	150,000	650,000
2010 年度	1,700,000	510,000	2,210,000
2011 年度	500,000	150,000	650,000
年度			
年度			
総　計	2,700,000	810,000	3,510,000

研究分野：工学
科研費の分科・細目：機械工学・知能機械学、機械システム
キーワード：機械知能、知能ロボット、制御工学

１．研究開始当初の背景

２足歩行ロボットが多くの研究施設・企業で開発されていたが、その行動の決定や歩容の決定については前もっての膨大な事前計算が必要であり、それらをロボットが自律的に決定出来るように知能化することが大き

な問題となっていた。また、長期歩行を実現するため、着地時の衝撃等からの破損を防止でき、衝撃エネルギーを吸収・回復してエネルギー効率を上げる目的で、足部に柔軟性を導入する必要性が認識されていた。

２．研究の目的

本研究の目的は、

(1) 実環境でロボットが自律的かつ効率的に最適行動を学習できる強化学習法の研究、
(2) その２足歩行ロボットへの応用、
(3) 関節剛性を調整可能な腱駆動２足歩行ロボットの機構と制御

である。

３．研究の方法

(1) については改良型罰回避政策形成アルゴリズムと報酬割り当て法を組み合わせた強化学習法を採用した。報酬割り当て法は学習速度が速い点でQ値学習法にまさるが、学習時間が長期にわたると、割引率の関係で初期の学習効率が著しく劣化する。そこで、改良型罰回避政策形成アルゴリズムを組み合わせた手法を用いた。このアルゴリズムは罰を負の報酬とするのではなく、失敗確率（罰ルール度）がある閾値γより、大きくなるとそのルールを罰上ルールとし、それ以後選択対象から強制的に排除するため、初期での学習効率を改善する事が期待される。その際、罰度閾値をどのように決定するかが不明であったので、学習によってこれを決定する手法を検討した。これはサッカーロボットを対象とし、４台のロボットのパスゲーム（Keep Away Task：３台でパスを回し、１台がそれを取りに行く。ボールを取られたり、ボールが領域内から出ると終了するゲーム）での最適行動を学習させた。その結果をシミュレーションおよび実機実験で検証した。

(2) 上記の手法を腱駆動２足歩行ロボットの腰軌道の生成に拡張した。その際、計算の容易な静的安定歩行の腰軌道から、動的安定歩行の腰軌道を学習する問題を考えた。この場合、①状態量（センサ入力）が離散値と連続値の両方を含み、かつ、高次元となる事、②遊脚接地時には状態が不連続に変化すること、③腰の変位量(x, y)（出力）が連続値であること、また、④学習が長時間にわたること、等が問題となる。そこで、①に対してはガウス基底関数を用いて状態を離散化することとする、その際、②に対処するため、連続的変化の場合は経験方向を強化するため、楕円体形基底関数を用い、不連続変化の場合は球型基底関数を用いることとした。③については変位量 x, y それぞれを３次多項式を用いて 15 個に離散化し、255 個（15×15）の行動を構成し、さらにそれを 30 個のサブ行動に分割して行動数を減少させ、また、３次スプライン関数を用いて加速度まで連続な量に変換した。④については３つの改良を行った。まず、Online 報酬割り当て法を導入して記憶容量の削減を図った。次に、学習のスケジューリングを行った。すなわち、学習目標歩数を定め、ある指定回数その歩数までの歩行が成功すると、歩行距離を更に半歩伸ばすようにした。最後に、すでに十分な報酬を得た状態を固定状態（学習済み状態）とし、それ以後、この状態では行動を決定論的に決定する事とした。この有効性を確認するため、次の２つの戦略を考え、評価した。(a)ある指定回数報酬を得ると、その状態を固定状態とする。(b) 目標歩数まで指定回数歩行が成功すると、その段階で報酬を得た状態を全て固定状態とする。なお、学習中に罰を受けた場合は、その状態から遡って３つ前までの固定状態を通常状態に戻すようにした。これらの効果をシミュレーションによって評価した。また、水平歩行での学習結果を初期値として階段昇降での学習に用いた場合の効果、関節剛性の学習への適用も検討した。

(3) 開発した腱駆動２足歩行ロボットは片脚６自由度で膝から下の３関節を６本の腱で駆動するようになっている。また、各腱には NST(非線形バネ要素)が装着され、腱の剛性範囲を拡大すると共に、腱張力のセンサとしても用いられている。歩行時には１本の腱に最大で 400N 程度の腱張力が加わるため、小さなプーリーを用いると容易に破断するため、末端形 NST を用いた。制御システムには２台の FPGA(Field Program- mable Gate Array)を用いた。腱張力制御精度を改善するために I-PD　制御法を適用した。

４．研究成果

(1) 閾値γは、そのルールの使用回数に対する罰を受けた回数の比として定義され、その値が閾値を越えるとそのルールは罰ルールと見なされ、それ以後の選択肢から排除される。そのため、この閾値が小さすぎると多くのルールが罰ルールとなり、かえって学習を阻害するおそれがある。図１に各γに対する、学習の進行度を示す。横軸は試行回数、縦軸は過去 100 試行毎の平均パス成功回数である。それまで、試行錯誤的に行ってきたが、その単峰性に着目してγの決定法を提案し、

シミュレーションと実機実験で検証した。実機実験による一例を図２に示す。縦軸は10試行毎の平均パス成功回数を示している。100 回の試行で学習が進んでいることが分かる。なお、割引率 0.8 閾値 0.6 を使用した。

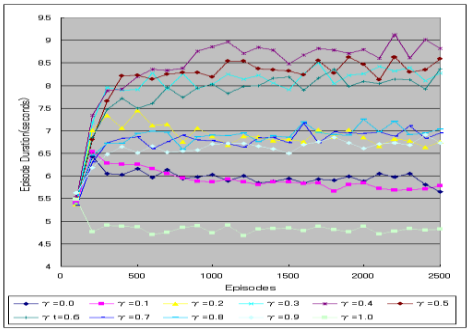


図１ γに対する学習の進度

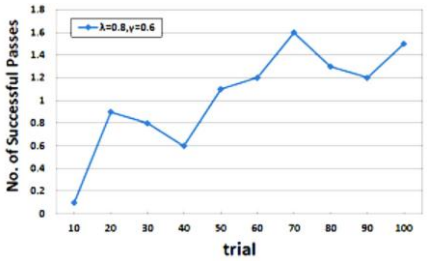


図２実機実験結果

(2) 上記の手法を２足歩行ロボットの腰軌道の学習に適用した。対象としたロボットを図３に示す。高さ約 100 cm、質量約 30 kg である。歩行は１歩 1.5sec で６歩あるいて停止する事とした。歩行開始から停止までの時間は 11.25 秒である。また、学習サンプリングタイムは 100msec とした。センサ入力は支持脚の指示値、支持様式の支持値（以上、離散値）、支持脚に対する腰の位置、腰の加速度、 ZMP 誤差（以上、連続値）とした。出力は腰座標の修正値（各 15 個の離散値）とし、３次スプライン関数を用いて補間した。罰として、実現不可能な姿勢、異常な関節速度、ZMP の支持多角形からの逸脱などを考えた。報酬としては関節角誤差、ZMP 誤差、消費エネルギーからなる評価関数を用い、目標歩数まで達したときの評価関数値が予想評価関数値よりも小さければ成功と考え、報酬 10,000 を与えた。小さい場合は報酬 100 を与えるが、成功回数には含まなかった。また、そのルールが７回以上選ばれ、その罰ルール度が 0.5 以上の場合は罰ルールとした。これらの学習用係数はGAを用いて最適化したものである。研究では、特に、前節(2)-④の戦略(a)と(b)の評価を行った。まず、各目標歩数に対してLT回成功すれば目標歩数を半歩増やすことにした。また、12 時間過ぎても学習が成功しない場合は学習失敗として打ち切った。以下の実験値は、各設定に対して１５回行った平均値を示している。戦略(a)に対する結果を図４に示す。ここでは LT=15 とし K 回報酬を得ればその状態を固定状態とすることとした。横軸は K の値であり、縦軸は平均学習時間[sec]（青色）と平均エピソード長（赤色）である。K=∞の時は固定状態を導入しない場合に相当する。これより、上手く K を設定することによって学習時間を半減できることが分かる。図５は戦略(b)に対する平均学習時間[sec]であり、青は固定状態を導入しない場合、赤は導入した場合を示している。LT が３以下で導入の効果が大きく現れている。特に LT=1 では平均約１時間で学習が終了しており、また、成功率も 100%であった。この手法を階段上り歩行にも適用し有効であることを確認にしている。また、関節剛性調整にも適用している。

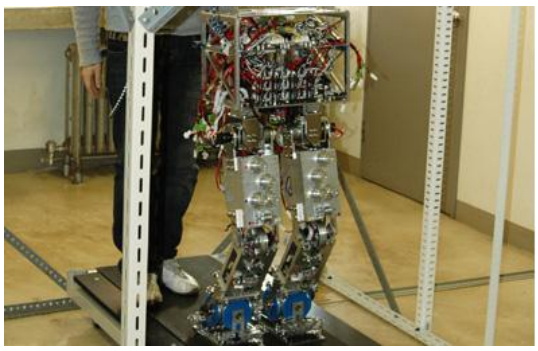


図３　腱駆動式２足歩行ロボット

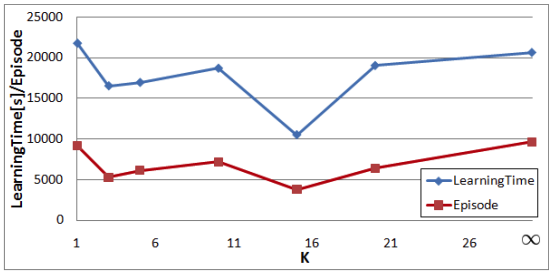


図４　戦略(a)

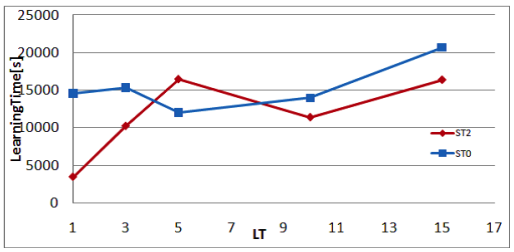


図５　戦略(b)

(3) 図３に示した腱駆動２足歩行ロボットの機構と制御について研究した。腱（ワイヤー）の剛性を調整可能にし、かつ大きな張力に耐えるように図６に示す NST を使用した。その特性を図７に示す。図７(a)は腱の伸びに対する腱張力であり、(b)は腱張力に対する等価剛性を示している。高い張力に耐え、剛性が腱張力にほぼ比例していることが分かる。図８に２つのFPGAを用いたコントローラを示す。大まかに言って、股部両足６関節の制御と両足腱駆動部の目標腱張力計算を mainFPGA で行い、両足腱駆動部の張力制御を subFPGA で行うことで、腱駆動部のサンプリングタイムを小さくしている。図９に使用した腱張力 I-PD 制御ループを示す。目標腱張力と誤差張力の積分値を公称腱張力とし、それに対応する腱の伸びを求め、目標モータ角を算出した後、モータ角のフィードファードDP制御を行っている。図１０は目標腱張力を 10-30-70-10N とステップ状に変化させたときの制御例である。

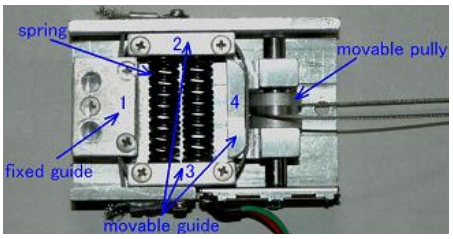
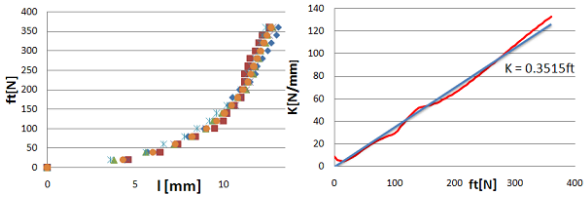


図６　mpNST



(a)　(b)

図７　NST の特性

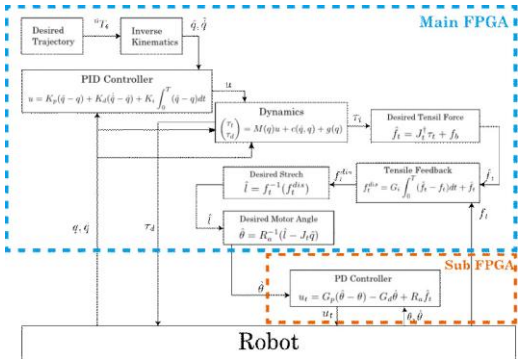


図８２つの FPGA を使ったコントローラ

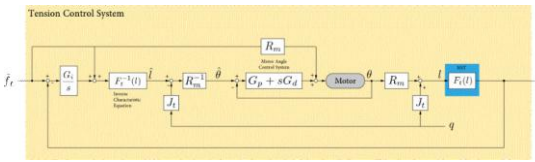


図９腱張力 I-PD 制御ループ

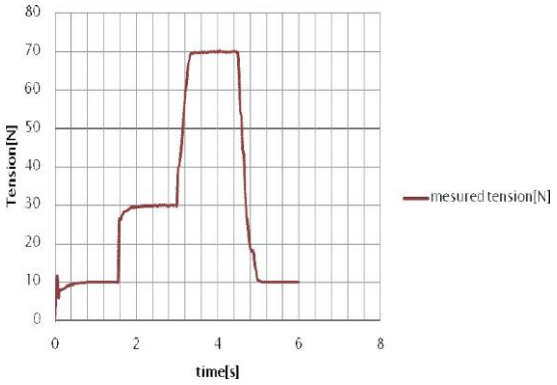


図１０　制御例

これらの成果は国内の学会や国際会議で発表され、特に、固定状態の導入に関しては高い評価を得た。今後の問題としては、罰伝播速度の改善、種々の歩容に対する腰軌道の学習と関節剛性の学習のスケジューリングなどがある。一定地点での屈伸運動から、順次、種々の動作を獲得していければと考えている。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 4 件）

(1) Kazuteru Miyazaki, Masaki Itou, and Hiroaki Kobayashi, Evaluation of the Improved Penalty Avoiding Rational Policy

Making Algorithm in Real World Environment, Lecture Notes in Computer Science, Vol. 7196, pp. 270-280, 2012,査読有り

(2) Seiya Kuroda, Kazuteru Miyazaki and Hiroaki Kobayashi, Introduction of Fixed Mode States into Online Profit Sharing and Its Application to Waist Trajectory Generation of Biped Robot, Lecture Notes in Computer Science, Vol. 7188, 2012, pp. 297-308、査読有り

(3) Kazuteru Miyazaki, Ryouhei Kobayashi, and Hiroaki Kobayashi, Threshold Learning in the Improved Penalty Avoiding Rational Policy Marking Algorithm, Proc. of SICE Annual Conference 2010, pp. 3240-3245, 2010、査読有り

(4) Takuji Watanabe, Kazuteru Miyazaki, and Hiroaki Kobayashi, A New Improved PenaltyAvoiding Rational Policy Making Algorithm for Keepaway with Conti-nuous State Spaces,Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.13,No. 6, pp. 675-683, 2009, 査読有り

〔学会発表〕(計 9 件)

(1) 伊藤大貴、岡島勇也、田中純夫、小林博明、宮崎和光、腱駆動２足歩行ロボットにおける腰軌道の強化学習への固定状態導入による効率化の研究、第 54 回自動制御連合講演会、2011 年 11 月 20 日、豊橋技術科学大学

(2) 村岡宏紀、宮崎和光、小林博明、罰と報酬を用いる強化学習の失敗確率伝播に関する研究、第 54 回自動制御連合講演会、2011 年 11 月 20 日、豊橋技術科学大学

(3) Seiya Kuroda, Kazuteru Miyazaki and Hiroaki Kobayashi, Introduction of Fixed Mode States into Online Profit Sharing and Its Application to Waist Trajectory Generation of Biped Robot, The 9th European Workshop on Reinforcement Learning (EWRL-9), The 9th European Workshop on Reinforcement Learning (EWRL-9), 2011 年 9 月 11 日, Athens Royal Olympic Hotel

(4) 黒田聖弥、日野雄太、岡島勇也、田中純夫、兵頭和人、小林博明、腱駆動２足歩行ロボットの開発と腰軌道および腱張力の強化学習―その２、第 53 回自動制御連合講演会、2010 年 11 月 4 日、高知市高知城ホール

(5) 伊藤昌樹、宮崎和光、小林博明、マルチエージェント連続タスクへの改良型罰回避政策形成アルゴリズムの適用とサッカーロボットを用いた実験による評価、第５３回自動制御連合講演会、2010 年 11 月 4 日、高知市高知城ホール

(6) Kazuteru Miyazaki, Ryouhei Kobayashi, and Hiroaki Kobayashi, Threshold Learning in the Improved Penalty Avoiding Rational Policy Marking Algorithm, SICE Annual Conference 2010, 2010 年 8 月 21 日, Grand Hotel,Taipei, Taiwan

(7) 黒田聖也，平野晃一郎，小林博明，田中純夫、腱駆動２足歩行ロボットの開発と腰軌道および腱張力の強化学習、日本機械学会関東支部第 16 期総会講演会、2010 年３月 10 日、明治大学アカデミーコモン

(8) 小林諒平，宮崎和光，小林博明、改良型罰回避政策形成アルゴリズムへの罰基底度決定機構の導入と評価、日本機械学会関東支部第 16 期総会講演会、2010 年３月 10 日、明治大学アカデミーコモン

(9) 小林諒平，宮崎和光,小林博明, 罰基底度閾値の学習機能を有する改良型罰回避政策形成アルゴリズムの提案, 第 52 回自動制御連合講演会, 2009 年 11 月 22 日,大阪大学基礎工学研究科

〔図書〕(計 0 件)

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
ホームページ等
http://www.messe.meiji.ac.jp/~iml/sub01/theme.html

６．研究組織
(1)研究代表者
小林　博明（KOBAYASHI HIROAKI）
明治大学・理工学部・教授
研究者番号：60130811

(2)研究分担者
田中　純夫（TANAKA SUMIO）
明治大学・理工学部・講師
研究者番号：40287884

(3)連携研究者
兵頭　和人（HYODO KAZUHITO）
神奈川工科大学・工学部・教授
研究者番号：10271371

宮崎　和光（MIYAZAKI KAZUTERU）
独立行政法人大学評価・学位授与機構・准教授
研究者番号：20282866