(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第4644833号
(P4644833)

(45)発行日　平成23年3月9日(2011.3.9)　(24)登録日　平成22年12月17日(2010.12.17)

(51)Int.Cl.　F I
B25J 5/00 (2006.01)　B25J 5/00 E
B25J 5/00 F

請求項の数 3　(全 34 頁)

(21)出願番号　特願2004-105291(P2004-105291)
(22)出願日　平成16年3月31日(2004.3.31)
(65)公開番号　特開2005-288594(P2005-288594A)
(43)公開日　平成17年10月20日(2005.10.20)
審査請求日　平成19年3月14日(2007.3.14)

(73)特許権者　393031586
株式会社国際電気通信基礎技術研究所
京都府相楽郡精華町光台二丁目２番地２
(74)代理人　110001195
特許業務法人深見特許事務所
(74)代理人　100064746
弁理士　深見　久郎
(74)代理人　100085132
弁理士　森田　俊雄
(74)代理人　100083703
弁理士　仲村　義平
(74)代理人　100096781
弁理士　堀井　豊
(74)代理人　100098316
弁理士　野田　久登

最終頁に続く

(54)【発明の名称】２足歩行移動装置

(57)【特許請求の範囲】
【請求項１】
右上腿部と前記右上腿部に右膝の関節を介して接続する右下腿部とを備える右脚と、
左上腿部と前記左上腿部に左膝の関節を介して接続する左下腿部とを備える左脚と、
前記右脚または前記左脚の接地を検出し、前記右上腿部に対して前記右下腿部が成す右膝角と前記左上腿部に対して前記左下腿部が成す左膝角とを検出するための第１のセンサ群とを備え、
前記右脚および前記左脚に接続し、前記右脚および前記左脚を駆動して２足歩行を行わせるための腰部と、
前記腰部のピッチ角を検出するためのピッチ角センサと、
前記腰部と前記右上腿部との成す右関節角と前記腰部と前記左上腿部との成す左関節角とを検出するための第２のセンサ群とをさらに備え、
前記腰部は、
前記右下腿部および前記左下腿部に対する状態マシンに基づいて、前記右下腿部および前記左下腿部を駆動する下腿制御信号を生成する下腿制御部と、
前記第２のセンサ群と前記ピッチ角センサの検知結果を受けて、前記右上腿部および前記左上腿部を駆動するための上腿制御信号を生成するための動的制御装置とを備え、
前記動的制御装置は、
前記右上腿部および前記左上腿部にそれぞれ対応する周期的な時間発展を行う内部状態を有するセントラルパターンジェネレータに対して、強化学習に基づいて前記右上腿部お

よび前記左上腿部の状態推定を行い、前記内部状態に対応する目標値に対するＰＤサーボ系の出力として、前記右上腿部および前記左上腿部に対する前記上腿制御信号を生成する制御手段を含み、

前記上腿制御信号および前記下腿制御信号に基づいて、前記右脚および前記左脚を駆動するための駆動手段をさらに備え、

前記下腿制御部の前記状態マシンは、

前記右脚および前記左脚のそれぞれについて、第１の膝屈曲状態と、第１の膝伸長状態と、第２の膝屈曲状態と、第２の膝伸長状態とを順次遷移する状態マシンであって、

前記第１の膝屈曲状態から前記第１の膝伸長状態への遷移および前記第２の膝屈曲状態から前記第２の膝伸長状態への遷移は、前記第２のセンサ群の検知結果に基づいて、前記左上腿部と前記右上腿部の成す角度が所定値を下回ることを条件として遷移が発生し、

前記第１の膝伸長状態から前記第２の膝屈曲状態への遷移および前記第２の膝伸長状態から前記第１の膝屈曲状態への遷移は、前記第１のセンサ群の検知結果に基づいて、前記左脚または前記右脚の接地に応じて発生し、

前記駆動手段は、前記状態マシンの状態に応じて予め定められた前記右膝角および前記左膝角についての目標角度と前記第１のセンサ群により検知される前記右膝角および前記左膝角とに基づくＰＤサーボ系の出力として、前記右膝および前記左膝とに加えるトルクを出力する膝駆動手段を含む、２足歩行移動装置。

【請求項２】

前記制御手段は、

前記右上腿部および前記左上腿部の状態情報に基づいて、価値関数と前記強化学習中に得られる報酬系列とに基づいて、フィードバックパラメータを更新するフィードバック制御器を含み、

前記フィードバック制御器からの出力に応じて変化する前記セントラルパターンジェネレータの前記内部状態に基づいて、前記上腿制御信号を生成する、請求項１記載の２足歩行移動装置。

【請求項３】

前記強化学習は、方策勾配法により行われる、請求項２記載の２足歩行移動装置。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、非線形の制御対象である２足歩行の周期運動を実現する上で有効となる動的制御装置を用いた２足歩行移動装置に関する。

【背景技術】

【０００２】

ロボットなどを制御しようとするとき、センサノイズや、そもそもセンサを配置することができないことによって、制御のために必要な状態変数を直接的には測定できない状況が考えられる。

【０００３】

そのような場合、状態観測器（オブザーバ）（たとえば、非特許文献１を参照）やカルマンフィルタ（たとえば、非特許文献２、非特許文献３を参照）を用いることが一般的である。しかしながら、対象のダイナミクスが非線形である場合、これらの手法では隠れ状態の推定は困難な場合がある。近年、非線形系に適用可能なオブザーバの提案がなされているものの、それぞれ特定の条件を満たさなければ適用できないなどの問題がある（たとえば、非特許文献４、非特許文献５を参照）。

【０００４】

また、拡張カルマンフィルタ(局所線形モデルを用いて状態分布の更新を行う)（たとえば、非特許文献３を参照）やモンテカルロフィルタ（モンテカルロ法により生成した多数の粒子により状態分布を近似する）（たとえば、非特許文献６を参照）などの手法が非線形系での状態推定法として知られているが、いずれも状態分布を陽に求めなければならな

い。

【０００５】

すなわち、従来の制御器は、基本的には、現在の状態を観測し、そこからの直接の写像によって制御出力を与える必要がある。

【非特許文献１】D.G. Luenberger著、“An introduction to observers”、IEEE Trans., AC, Vol.16, pp. 596--602, 1971.

【非特許文献２】R.E. Kalman and R.S. Bucy著、“New results in linear filtering and prediction theory”、Trans., ASME, Series D, J. of Basic Engineering, Vol.83, No.1, pp. 95--108, 1961.

【非特許文献３】F.L. Lewis著、“Optimal Estimation: with an Introduction to Stochastic Control Theory”、John Wiley \& Sons, 1977.

【非特許文献４】志水清孝, 鈴木俊輔, 田中哲史著、“こう配降下法による非線形オブザーバ（非線形システムの状態観測器）”、電子情報通信学会論文誌 A, Vol. J83-A, No.8, pp. 956--964, 2000.

【非特許文献５】H.Nijimeijer and T.L. Fossen著、“Directions in Nonlinear Observer Design”、Springer-Verlag, London, 1999.

【非特許文献６】G.Kitagawa著、“Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Models”、Journal of Computational and Graphical Statistics, Vol.5, pp. 1--25, 1996.

【発明の開示】

【発明が解決しようとする課題】

【０００６】

本発明の目的は、制御器に内部状態とダイナミクスを持たせることで、周期運動に対し状態観測器の学習を容易とすることが可能な動的制御装置を用いた２足歩行移動装置を提供することである。

【課題を解決するための手段】

【０００７】

本発明では、学習の目的は、ある瞬間での推定誤差を少なくするのではなく、タスクを行っている期間を通じての推定誤差を少なくすることである。そこで、方策勾配法（強化学習）の枠組みを用いて非線形状態観測器の構築を行っている。

【０００８】

本発明の１つの局面では、状態観測器のダイナミクスへの入力を学習器の行動と考え、現在の観測可能な状態（観測出力）、状態観測器の状態、制御器の出力から、適切に状態観測器の状態を制御対象の状態へと導く行動則を方策勾配法によって獲得している。

【０００９】

したがって、この発明の１つの局面に従うと、２足歩行移動装置であって、右上腿部と右上腿部に右膝の関節を介して接続する右下腿部とを備える右脚と、左上腿部と左上腿部に左膝の関節を介して接続する左下腿部とを備える左脚と、右脚または左脚の接地を検出し、右上腿部に対して右下腿部が成す右膝角と左上腿部に対して左下腿部が成す左膝角とを検出するための第１のセンサ群と、右脚および左脚に接続し、右脚および左脚を駆動して２足歩行を行わせるための腰部と、腰部のピッチ角を検出するためのピッチ角センサと、腰部と右上腿部との成す右関節角と腰部と左上腿部との成す左関節角とを検出するための第２のセンサ群とをさらに備え、腰部は、右下腿部および左下腿部に対する状態マシンに基づいて、右下腿部および左下腿部を駆動する下腿制御信号を生成する下腿制御部と、第２のセンサ群とピッチ角センサの検知結果を受けて、右上腿部および左上腿部を駆動するための上腿制御信号を生成するための動的制御装置とを備え、前記動的制御装置は、前記右上腿部および前記左上腿部にそれぞれ対応する周期的な時間発展を行う内部状態を有するセントラルパターンジェネレータに対して、強化学習に基づいて右上腿部および左上腿部の状態推定を行い、内部状態に対応する目標値に対するＰＤサーボ系の出力として、右上腿部および左上腿部に対する上腿制御信号を生成する制御手段を含み、上腿制御信号

および下腿制御信号に基づいて、右脚および左脚を駆動するための駆動手段をさらに備え、下腿制御部の状態マシンは、右脚および左脚のそれぞれについて、第１の膝屈曲状態と、第１の膝伸長状態と、第２の膝屈曲状態と、第２の膝伸長状態とを順次遷移する状態マシンであって、第１の膝屈曲状態から第１の膝伸長状態への遷移および第２の膝屈曲状態から第２の膝伸長状態への遷移は、第２のセンサ群の検知結果に基づいて、左上腿部と右上腿部の成す角度が所定値を下回ることを条件として遷移が発生し、第１の膝伸長状態から第２の膝屈曲状態への遷移および第２の膝伸長状態から第１の膝屈曲状態への遷移は、第１のセンサ群の検知結果に基づいて、左脚または右脚の接地に応じて発生し、駆動手段は、状態マシンの状態に応じて予め定められた右膝角および左膝角についての目標角度と第１のセンサ群により検知される右膝角および左膝角とに基づくＰＤサーボ系の出力として、右膝および左膝とに加えるトルクを出力する膝駆動手段を含む。

【００１０】

好ましくは、制御手段は、右上腿部および左上腿部の状態情報に基づいて、価値関数と強化学習中に得られる報酬系列とに基づいて、フィードバックパラメータを更新するフィードバック制御器を含み、フィードバック制御器からの出力に応じて変化するセントラルパターンジェネレータの内部状態に基づいて、上腿制御信号を生成する。

【００１１】

好ましくは、強化学習は、方策勾配法により行われる。

【発明の効果】

【００１５】

本発明では、行動則自体が内部変数とその微分方程式によって表されるダイナミクスを持つため、行動則と物理系の引き込みの性質を利用することができ、周期運動に対して状態推定を行う場合に、状態観測器の学習を容易とすることが可能である。さらに、センサ入力のノイズや時間遅れに対しても、ロバストな性質を持つ制御器を実現することができる。

【００１６】

さらに、本発明によれば、２足歩行移動装置の自由度が増加し、状態空間が高次元となった場合でも、周期運動に対し状態観測器の学習を容易とすることが可能な動的制御装置動的制御装置を用いた２足歩行移動装置を提供することができる。

【発明を実施するための最良の形態】

【００１７】

以下、図面を参照して本発明の実施の形態について説明する。

【００１８】

以下の説明では、一例として、本発明を２足歩行の制御に適用する場合を説明するが、本発明は、必ずしもこのような場合に限定されるものではなく、たとえば、より一般的に周期運動を行う系に対して有効な制御システムを提供するものである。特に、本発明は、周期運動を行う劣駆動（機械）系に適用するのに適した制御システムを提供する。

【００１９】

［実施の形態１］

（本発明のシステム構成）

図１は、本発明の動的制御装置を用いた２足歩行移動システム１０００の一例を示す概念図である。

【００２０】

図１を参照して、システム１０００は、動的制御装置１００と、動的制御装置１００の上部に設けられる胴部４０と、動的制御装置１００により駆動制御される脚部とを備える。脚部は、右脚１０ｒと左脚１０ｌとを有し、各脚は、接地面近傍に設けられるセンサ２０ｒおよび２０ｌとを備える。一方、胴部４０には、センサ３０が設けられる。

【００２１】

センサ２０ｒは、胴部４０の中心線４に対する右脚の角度θ_r、角速度$d\theta_r/dt$という情報を検出し、また、センサ２０ｌは、中心線４に対する左脚の角度θ_l、角速度$d\theta_l$

/ｄｔという情報を検出し、それぞれ、動的制御装置１００に通知する。さらに、センサ３０は、鉛直方向２に対する胴部４０のピッチ角θ_p、角速度$d\theta_p/dt$という情報を検出し、それぞれ、動的制御装置１００に通知する。

【００２２】

動的制御装置１００は、センサ２０ｒ、センサ２０ｌ、センサ３０からの情報に基づいて、右脚１０ｒおよび左脚１０ｌの動作を制御する。

【００２３】

図２は、図１に示した動的制御装置１００の構成を示すブロック図である。

【００２４】

図２を参照して、動的制御装置１００は、センサ２０ｒ、センサ２０ｌ、センサ３０からの信号を受け取る通信インタフェース１０６と、後に説明する制御パラメータやセンサからの情報を格納しておくための記憶装置１０４と、センサ２０ｒ、センサ２０ｌ、センサ３０からの情報を用いて学習して獲得した動的行動則に基づき、制御信号を生成する演算処理部１０２と、演算処理部１０２からの制御信号に基づいて、右脚１０ｒおよび左脚１０ｌの駆動制御を行うための駆動部１０８とを備える。

【００２５】

以下では、動的制御装置１００の制御動作のための準備の処理および制御動作について説明する。

【００２６】

（１－１．動的行動則）

まず、本発明の制御動作を説明する前提として、「動的行動則」について説明する。

【００２７】

図３は、このような動的行動則を説明するための概念図である。

【００２８】

「動的行動則」とは、図３（１）に示すような行動則自体が内部変数とその微分方程式によって表されるダイナミクスを持つ枠組である。

【００２９】

これをより具体的に表現すると、図３（２）に示すように観測器が内部変数及びその微分方程式を持つようなものや、図３（３）に示すような制御器が内部変数及びその微分方程式を持つようなものである場合が考えられる。

【００３０】

このような「動的行動則」に基づいて、制御対象を制御するような制御装置を「動的制御装置」と呼ぶことにする。

【００３１】

以下では、まず、図３（３）の枠組を用いることで、動的制御装置１００により、２足歩行運動を実現する場合を考える。

【００３２】

行動則が内部状態を持つことによって、行動則と物理系の引き込みの性質を利用することが出来るため、周期運動や状態推定を行う場合に有効であると考えられる。さらに、センサ入力のノイズや時間遅れに対しても、ある程度ロバストな性質を持つことが期待できる。

【００３３】

このような行動則を学習によって獲得する場合、行動則の内部状態が隠れ変数となることは問題となる。しかし、行動則の内部状態が物理系の状態に対して引き込むことで、それぞれの状態は一意に対応するようになる。ただし、過渡的な状態では隠れ状態を扱う必要がある。

【００３４】

そこで本発明では、動的行動則の獲得手法として、後に説明するように、隠れ変数が存在する環境においても適用可能な方策勾配法を用いる。

【００３５】

以下では、本発明の学習システムおよび、動的行動則を構成するセントラルパターンジェネレータ（central pattern generator：ＣＰＧ)とＣＰＧへのフィードバック制御器について説明を行う。

【００３６】

（１－２．学習システム）

以下の説明では、周期運動の例として、３リンク２足歩行ロボットモデルを用いた２足歩行運動に対して動的行動則を適用する。

【００３７】

図４は、図２で説明した演算処理部１０２の行う処理を示す機能ブロック図である。以下に説明するとおり、演算処理部１０２は、学習システムとして機能する。

【００３８】

図４に示すとおり、この学習システムは、基本的に、ＣＰＧ処理部１０２６とフィードバック制御器１０２２によって動的行動則を構成する。学習に用いる状態ｘは、以下の式で表される。

【００３９】

【数１】

$$x=(x_1,x_2,x_3,x_4)^T=(\theta_l+\theta_p,\theta_r+\theta_p,\dot{\theta}_l+\dot{\theta}_p,\dot{\theta}_r+\dot{\theta}_p)^T$$

【００４０】

ただし、上述のとおり、θ_r、θ_lは、それぞれロボットの鉛直方向からの左右の脚１０ｒ、１０ｌの角度であり、θ_pは胴体４０のピッチ角である。

【００４１】

つまり、学習システムはロボットから直接得られる信号のみで状態空間を構成しており、ＣＰＧの内部状態を用いていないという特徴を有している。

【００４２】

また、ここでは全状態観測を仮定し、ｙ＝ｘであるとする。

【００４３】

（１－３．セントラルパターンジェネレータ（ＣＰＧ））

演算処理部１０２により実現される学習システムで、動的行動則を構成するＣＰＧ処理部１０２６の構成として、以下の式で表される神経振動子モデルを用いる。なお、このような神経振動子モデルについては、たとえば、文献：Kiyoshi Matsuoka著、“Sustained oscillations generated by mutually inhibiting neurons with adaptation.”、Biological Cybernetics, Vol.52, pp. 367-376, 1985に開示がある。

【００４４】

【数２】

$$\tau \dot{z}_1 = -z_1 - \omega_{12} q_2 - \omega_{13} q_3 - \beta p_1 + z_0 + a_1 \quad (1)$$

$$\tau' \dot{p}_1 = -p_1 + q_1 \quad (2)$$

$$q_1 = max(0, z_1) \quad (3)$$

$$\tau \dot{z}_2 = -z_2 - \omega_{21} q_1 - \omega_{24} q_4 - \beta p_2 + z_0 + a_2 \quad (4)$$

$$\tau' \dot{p}_2 = -p_2 + q_2 \quad (5)$$

$$q_2 = max(0, z_2) \quad (6)$$

$$\tau \dot{z}_3 = -z_3 - \omega_{34} q_4 - \omega_{31} q_1 - \beta p_3 + z_0 + a_3 \quad (7)$$

$$\tau' \dot{p}_3 = -p_3 + q_3 \quad (8)$$

$$q_3 = max(0, z_3) \quad (9)$$

$$\tau \dot{z}_4 = -z_4 - \omega_{43} q_3 - \omega_{42} q_2 - \beta p_4 + z_0 + a_4 \quad (10)$$

$$\tau' \dot{p}_4 = -p_4 + q_4 \quad (11)$$

$$q_4 = max(0, z_4) \quad (12)$$

【００４５】

ここで、変数：ｚ，ｐはニューロン内の状態、ｑはニューロンの出力、z_0は持続入力、定数βはニューロンの疲労係数、τ、τ´は、ｚ，ｐの時定数、ωは拮抗ニューロン間の結合係数である。また、ａは、後に説明するフィードバック制御器からの出力項である。

【００４６】

図５は、式（１）～（１２）で表される神経振動子モデルによるＣＰＧを示す概念図である。図５においては、ニューロン内の状態ｚ、ｐの間で、相互に正の結合を行うものは白丸で、負の結合を行うものは黒丸で示している。

【００４７】

図６は、このようなＣＰＧの出力ｑを構成する変数z_1、z_2の波形を示す図である。なお、この計算では、例として、τ＝０．０５、τ´＝０．６、β＝２．５、ω＝２．０、z_0＝０．１、ａ＝０を用いた。

【００４８】

式（１）～（１２）で表されるモデルにしたがって、ＣＰＧの内部状態変数z_1、z_2が周期的に変化していることがわかる。

【００４９】

さらに、図４の学習システムのＰＤサーボ処理部１０２８では、以下に示すとおり、各ニューロンの出力の差を両脚のサーボ系の目標関節角θ^dとした。

【００５０】

【数３】

$$\theta_l^d = (q_1 - q_2) \quad (13)$$

$$\theta_r^d = (q_3 - q_4) \quad (14)$$

【００５１】
ただし、θ_l^dは左脚の目標関節角、θ_r^dは右脚の目標関節角である。
【００５２】
ＰＤサーボ処理部１０２８の結果出力されるロボットへのトルク入力ｕは、次に示すＰＤサーボ系の出力を用いる。
【００５３】
【数４】

$$u_l = K_p(\theta_l^d - \theta_l) - K_d\dot{\theta}_l \qquad (15)$$

$$u_r = K_p(\theta_r^d - \theta_r) - K_d\dot{\theta}_r \qquad (16)$$

【００５４】
ただし、u_lは左脚に対するトルク入力、u_rは右脚に対するトルク入力である。また、K_pは位置ゲイン、K_dは速度ゲインである。
【００５５】
（１－４．フィードバック制御器１０２２）
上述のＣＰＧへのフィードバック制御器１０２２は、次の確率分布（１７）によって表される。
【００５６】
【数５】

$$\pi(v_j, \mathbf{x}; \mathbf{w}) = \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left(\frac{-(v_j - \mu_j)^2}{2\sigma_j^2}\right) \qquad (17)$$

【００５７】
ただし、ｘは制御対象の状態ベクトル、ｗはパラメータベクトルである。
従ってｊ番目の出力の実現値v_jは、以下の式（１８）によって与えられる。
【００５８】
【数６】

$$v_j(t) = \mu_j(t) + \sigma_j n_j(t) \qquad (18)$$

【００５９】
ただし、ｎｊ（ｔ）～Ｎ（０，１）であり、Ｎ（０，１）は平均０、分散１の正規分布を表す。
【００６０】
ここでは出力を飽和させるために、出力飽和処理部１０２４において、関数ｄ（）を用いて以下の式（１９）のように、最終的な制御器の出力a_j（ｔ）を決定する。
【００６１】
【数７】

$$a_j(t) = a_j^{max} d(v_j(t)) \qquad (19)$$

【００６２】
ただし、以下の説明では、一例として、ｄ（）としては、以下の式を用いる。
【００６３】

【数８】

$$d(x)=\frac{2}{\pi}\arctan(\frac{\pi}{2}x)$$

【００６４】

ここでのａ$_j$(j=１～４)は、式（１）の左右の脚の神経振動子の伸筋、屈筋にそれぞれ対応する。

【００６５】

（２．方策勾配法）

「方策勾配法」とは、パラメータ化された確率的方策に従って行動選択を行い、方策を改善する方向に方策のパラメータを少しずつ更新する強化学習手法の１種である。以下に方策勾配法を用いた行動則の学習方法について述べる。

【００６６】

（２－１．連続時間・状態系でのテンポラル・ディファレンス（Temporal Difference）誤差）

連続時間・状態系のダイナミクスを以下の式（２０）で表す。

【００６７】

【数９】

$$\frac{d\mathbf{x}(t)}{dt}=f(\mathbf{x}(t),\mathbf{u}(t)) \qquad (20)$$

【００６８】

ただし、ｘ∈Ｘ⊂Ｒ^{n}は状態、ｕ∈Ｕ⊂Ｒ^{m}は制御入力を表す。

【００６９】

報酬は状態と制御入力の関数として、以下の式（２１）で与えられるとする。

【００７０】

【数１０】

$$r(t)=r(\mathbf{x}(t),\mathbf{u}(t)) \qquad (21)$$

【００７１】

ある制御則π（ｕ（ｔ）｜ｘ（ｔ））のもとで、状態ｘ（ｔ）の価値関数を以下の式（２２）で定義する。

【００７２】

【数１１】

$$V^{\pi}(\mathbf{x}(t))=E\left\{\int_t^{\infty}e^{-\frac{s-t}{\tau}}r(\mathbf{x}(s),\mathbf{u}(s))\,ds \mid \pi\right\} \qquad (22)$$

【００７３】

ただし、τは価値関数の時定数である。また、式（２２）の両辺の時間微分から、以下の式（２３）という拘束条件が与えられる。

【００７４】

【数１２】

$$\frac{dV^{\pi}(\mathbf{x}(t))}{dt}=\frac{1}{\tau}V^{\pi}(\mathbf{x}(t))-r(t) \qquad (23)$$

【００７５】
Ｖ（ｘ（ｔ））＝Ｖ（ｘ（ｔ）；ｗ）を価値関数の予測値とする。ただし、ｗは評価値の予測値のパラメータである。
【００７６】
予測が正しければ、式（２３）を満たす。予測が正しくない場合、下式（２４）に示した予測誤差を減らすように学習を行う。
【００７７】
【数１３】

$$\delta(t) = r(t) - \frac{1}{\tau}V(t) + \dot{V}(t) \qquad (24)$$

【００７８】
上式は連続時間系でのＴＤ誤差である。
【００７９】
（２－２．方策勾配法の一般論）
動的計画法やグリーディ方策（greedy policy）などの価値関数の評価を基に学習を行う場合では、環境がマルコフ決定過程である必要があるが、実問題に適用する場合には、ノイズやセンサの能力によってマルコフ決定過程を保証することは困難である。しかし、方策勾配法は、価値関数と共に、試行中に得られた累積報酬系列を考慮することで、環境が非マルコフ決定過程(POMDP)でも適用することが出来る。
【００８０】
ここで、パラメータｗを持つ方策πwを用いた場合、以下の式（２５）が成り立つ。
【００８１】
【数１４】

$$\frac{\partial}{\partial w_i}E\{\, V(t) \mid \pi_{\mathbf{w}}\} = E\{\delta(t)e_i(t)\} \qquad (25)$$

【００８２】
ただし、以下の式（２６）が成り立つ。
【００８３】
【数１５】

$$\dot{e}_i(t) = -\frac{1}{\kappa}e_i(t) + \frac{\partial \ln \pi_{\mathbf{w}}}{\partial w_i} \qquad (26)$$

【００８４】
ここで、κはエリジビリティ・トレース（eligibility trace）の時定数である。テンポラル・ディファレンス誤差δと方策のエリジビリティ・トレースｅ（ｔ）により、価値関数の方策パラメータｗに関する勾配の不偏推定量を求めることが出来ることが与えられている。このような方策勾配法については、たとえば、文献：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-不完全なvalue-functionのもとでの強化学習”、人工知能学会誌，Vol.15，No.2，pp. 267-275，2000に記載がある。
【００８５】
よって、パラメータの更新則は次の式（２７）のようになる。
【００８６】

【数１６】

$$\dot{w}_i = \eta\delta(t)e(t) \qquad (27)$$

【００８７】

ただし、ηは学習率である。

【００８８】

（２－３．動的行動則の学習）

以下では、上述の方策勾配法を用いて、動的行動則の獲得を行う。

【００８９】

ここでは、式(１７)に示したフィードバック制御器の学習を行うことで望みの動的行動則を獲得することを考える。

【００９０】

（２－３－１．価値関数の更新）

まず、価値関数処理部１０３２において演算される、連続状態における価値関数の表現方法として、以下の式（２８）による正規化ガウス関数ネットワーク(normalized Gaussian network: ＮＧｎｅｔ)を用いる。なお、正規化ガウス関数ネットワークについては、後に説明する。

【００９１】

【数１７】

$$V(\mathbf{x}(t)) = \sum_i w_i^c b_i^c(\mathbf{x}(t)) \qquad (28)$$

【００９２】

ただし、b_i^c（）は、正規化処理部１０３０においてｘに施される基底関数であり、w_i^cは価値関数のパラメータである。

【００９３】

パラメータw_i^cに対するエリジビリティ・トレースe_i^cと、ＴＤ誤差を用いたパラメータw_i^cの更新式は、それぞれ以下の式（２９）および（３０）のようになる。

【００９４】

【数１８】

$$\dot{e}_i^c(t) = -\frac{1}{\kappa^c}e_i^c(t) + b_i^c(\mathbf{x}(t)) \qquad (29)$$

$$\dot{w}_i^c(t) = \alpha\delta(t)e_i^c(t) \qquad (30)$$

【００９５】

ただし、αは価値関数の学習率、κ^cはエリジビリティ・トレースの時定数である。

【００９６】

（２－３－２．フィードバック制御器の更新）

式(１７)に示した確率的なフィードバック制御器１０２２を用いる場合、そのｊ番目の出力の平均μ_jと標準偏差σ_jに関するエリジビリティは式(２６)右辺第２項と同様、それぞれ以下のように与えられる。

【００９７】

【数１９】

$$\frac{\partial \ln \pi}{\partial \mu_j} = \frac{a_t - \mu_j}{\sigma_j^2} \tag{31}$$

$$\frac{\partial \ln \pi}{\partial \sigma_j} = \frac{(a_t - \mu_j)^2 - \sigma_j^2}{\sigma_j^3} \tag{32}$$

【００９８】

ここではさらに、以下の式（３３）および（３４）のように、平均μを正規化ガウス関数ネットワークによって表し、標準偏差σをシグモイド関数によって表す。

【００９９】

【数２０】

$$\mu_j = \sum_i w_{ij}^{\mu} b_i^{\mu}(\mathbf{x}(t)) \tag{33}$$

$$\sigma_j = \frac{1}{1 + \exp(-w_j^{\sigma})} \tag{34}$$

【０１００】

ただし、ノーテーションとしては以下のとおりである。

【０１０１】

【数２１】

b_i^{μ} ：基底関数、

$\omega_{ij}^{\mu}, \omega_j^{\sigma}$ ：式（１７）に示したフィードバック制御器のパラメータ

【０１０２】

これらのパラメータに対応するエリジビリティは以下の式（３５）および（３６）のように求められる。

【０１０３】

【数２２】

$$\frac{\partial \ln \pi}{\partial w_{ij}^{\mu}} = \frac{\partial \ln \pi}{\partial \mu_j} \frac{\partial \mu_j}{\partial w_{ij}^{\mu}} = \frac{(a_t - \mu_j) b_i^{\mu}(\mathbf{x}(t))}{\sigma_j^2} \tag{35}$$

$$\frac{\partial \ln \pi}{\partial w_j^{\sigma}} = \frac{\partial \ln \pi}{\partial \sigma_j} \frac{\partial \sigma_j}{\partial w_j^{\sigma}} = \frac{((a_t - \mu_j)^2 - \sigma_j^2)(1 - \sigma_j)}{\sigma_j^2} \tag{36}$$

【０１０４】

上式(３５)(３６)と式(２６)(２７)を考慮すると、以下の式（３７）および（３８）のようなフィードバックパラメータの更新則が得られる。

【０１０５】

【数２３】

$$\dot{w}_{ij}^{\mu} = \beta^{\mu}\delta(t)e_{ij}^{\mu}(t) \qquad (37)$$

$$\dot{w}_{j}^{\sigma} = \beta^{\sigma}\delta(t)e_{j}^{\sigma}(t) \qquad (38)$$

【０１０６】
ただし、ノーテーションとしては以下のとおりである。

【０１０７】
【数２４】

$\beta^{\mu},\beta^{\sigma}$　　：学習率
$e_{ij}^{\mu}(t),e_{j}^{\sigma}(t)$：それぞれのパラメータのエリジビリティ・トレース

【０１０８】
また、式(３５)(３６)において、パラメータσが分母となっていることにより、σが０へと近付くとエリジビリティが発散することが問題となる。そこでエリジビリティ・トレースの更新には式(２６)の代わりに次式を用いる。

【０１０９】
【数２５】

$$\dot{e}_{ij}^{\mu}(t) = -\frac{1}{\kappa^{\mu}}e_{ij}(t) + \sigma_{j}^{2}\frac{\partial \pi_{\mathbf{w}}}{\partial w_{ij}} \qquad (39)$$

$$\dot{e}_{j}^{\sigma}(t) = -\frac{1}{\kappa^{\sigma}}e_{j}(t) + \sigma_{j}^{2}\frac{\partial \pi_{\mathbf{w}}}{\partial w_{j}} \qquad (40)$$

【０１１０】
ただし、ノーテーションとしては以下のとおりである。

【０１１１】
【数２６】

$\kappa^{\mu},\kappa^{\sigma}$：それぞれのパラメータのエリジビリティ・トレースの時定数

【０１１２】
（２－４．具体例）
図４に示した学習システムにおいて、数値シミュレーションを行った結果について以下説明する。

【０１１３】
このシミュレーションにおいて、図１に示した２足歩行移動システム１０００（２足歩行ロボット）は、脚長が０．２ｍ、両脚の質量がそれぞれ０．５ｋｇとし、胴体が０．１ｋｇであるものとした。さらに、膝関節がないことを考慮して、遊脚を振り出す場合は足先が地面を通過出来るように設定した。

【０１１４】
それぞれの学習パラメータは、以下のとおりである。

【０１１５】

【数２７】

学習率 ：$\alpha = 0.04$
$\beta^{\mu} = 0.2$
$\beta^{\sigma} = 0.01$
各エリジビリティの時定数： $\kappa = 2.0$
サーボ系のゲイン ： $K_p = 5.0$
$K_d = 0.1$

【０１１６】
また、ＮＧｎｅｔの基底関数は、実際にロボットが歩行運動を行う際に必要であると予想される状態空間に格子状に均等に配置することを考え、以下のようにする。

【０１１７】
【数２８】

$$x = (x_1, x_2, x_3, x_4)^T$$
$$= (\theta_l + \theta_p, \theta_r + \theta_p, \dot{\theta}_l + \dot{\theta}_p, \dot{\theta}_r + \dot{\theta}_p)^T \quad : (12, 6, 12, 6)$$

【０１１８】
この結果、計５１８４（＝１２×６×１２×６）個をそれぞれ、以下の範囲に均等に配置した。

【０１１９】
【数２９】

$$(-\frac{\pi}{2} \le x_1 \le \frac{\pi}{2}, -3\pi \le x_2 \le 3\pi, -\frac{\pi}{2} \le x_3 \le \frac{\pi}{2}, -3\pi \le x_4 \le 3\pi)$$

【０１２０】
報酬関数は以下の式で表す。

【０１２１】
【数３０】

$$r(\mathbf{x}) = k_H r_H(\mathbf{x}) + k_S r_S(\mathbf{x}) \qquad (41)$$

【０１２２】
ただし、それぞれがロボットの腰の高さに関する項r_H（ｔ）、歩行速度に関する項r_S（ｔ）は、以下の式で表される。

【０１２３】
【数３１】

$$r_H(x) = h_1 - h' - \min(f_l, f_r),$$
$$r_S = \max(-1, \min(\dot{x}_1, 1))$$

【０１２４】
ここで、h_1はロボットの腰の高さ、ｈ´は腰の高さのオフセット、f_l, f_rは左右の脚の高さである。したがって、式（４１）の右辺第１項は、ロボットの位置エネルギーに関連する量であり、右辺第２項はロボットの運動エネルギーに関連する量である。

【０１２５】
以下に説明するシミュレーションでは各パラメータは、$k_s = 0.06$、$k_H = 0.5$、

ｈ′＝０．１５とした。また、ＣＰＧのパラメータは（１－２．学習システム）で述べたものを用いている。

【０１２６】

計算機上でのロボット及びＣＰＧのダイナミクスの時間刻みは１ｍｓｅｃ、学習システムの時間刻みは１０ｍｓｅｃとした。

【０１２７】

また、シミュレーションにおいて、１学習試行の終了条件は以下のようにした。

【０１２８】

ｉ）１７７００ｍｓｅｃ経過(約１００歩の歩行終了後)

ｉｉ）転倒時(ただし、同時にｒ＝－１の報酬を与える)

（２－５．平地歩行の獲得及び、環境変化に対するロバスト性）

図７は、１試行で獲得した報酬の総和を、試行回数ごとに取った学習曲線を示す図である。図７においては、地面の傾斜０°のときの学習曲線を示している。

【０１２９】

図７より、学習は約３５０回で収束しており、定常歩行運動を獲得出来ていることが分かる。

【０１３０】

図８は、図７の学習曲線に対応する歩行の軌跡を示す図である。

【０１３１】

図８において、(1)は学習前の歩行軌跡、(2)は６００回学習後の歩行軌跡を示す。６００回の学習後では、歩幅が大きくなり歩行速度も向上して、良好な歩行軌跡が得られていることがわかる。

【０１３２】

また、６００回学習試行を行うことによって学習した各学習パラメータを用い、数度の傾斜を付けることによって環境を変化させた場合でも、ある程度歩行動作を維持することが可能である。さらに、数回の学習試行を行うことによって、新しい環境に適応することが出来る。これは、行動則の内部状態(ここではＣＰＧの内部状態)と、ロボットの状態が引き込みを行うことのよって、ロバストなリミットサイクルを構成しているからであると考えられる。

【０１３３】

図９は、図７で獲得した歩行において、ＣＰＧの内部状態とロボットの状態の間のリミットサイクルを、ＣＰＧの内部状態z_1と脚角度の時間変化として示す図である。

【０１３４】

また、図１０は、図７で獲得した歩行において、ＣＰＧの内部状態とロボットの状態の間のリミットサイクルを、脚角度、脚の角速度、ＣＰＧの内部状態z_1の関係として示す図である。

【０１３５】

外部からの擾乱に対しても、本発明の制御システムは、周期運動を継続させることが可能なことがわかる。

【０１３６】

（２－６．報酬と獲得した運動の関係）

式(４１)の報酬関数中の、速度項係数k_sを変化させた場合の、ロボットの歩行速度の関係を表１に示す。

【０１３７】

【表１】

表 1:　報酬の速度項係数と歩行速度との関係

報酬の速度項係数 k_s	ロボットの歩行速度 (m/sec)
0.06	0.345
0.1	0.388
0.6	0.419

【０１３８】

ここで、ロボットの腰の高さに関する項の係数は、前節と同様k_H＝０．５とした。

【０１３９】

表１より、速度項を増加させるとロボットの歩行速度も増加することが分かり、よってロボットのダイナミクスから構成するようなコントローラを陽に用いることなく、学習の報酬を変化させることによって、ロボットを制御出来ることが確認出来る。

【０１４０】

（２－７．センサノイズ・時間遅れに対するロバスト性）

図７で獲得された歩行を教師信号として学習した各パラメータを初期値として用い、さらに図４の学習システムからＣＰＧを取り除いたものを用いて、１５０回学習試行を行うことによって、内部状態を持たない行動則によって２足歩行運動を獲得した。これと、図７の学習によって獲得した歩行運動を用いて、コントローラのセンサノイズ及び時間遅れに対するロバスト性について比較を行った。

【０１４１】

センサノイズはx_1、x_3に対しては、Ｎ（０，０．０１）、x_2、x_4に対しては、Ｎ（０，０．０９）を用い、時間遅れは２０ｍｓｅｃとしてシミュレーションを行った。

【０１４２】

図１１は、センサノイズ・時間遅れに対するシミュレーション結果を示す図である。図１１において、（１）はＣＰＧ有り、（２）はＣＰＧ無しのコントローラで構成された歩行を示す。また、図１１において、（ａ）は通常の条件での歩行、（ｂ）はセンサノイズのある状態での歩行、（ｃ）は時間遅れがある場合の歩行であり，また、図１１中で、“→”はロボットの進行方向を表している。

【０１４３】

ＣＰＧを持たない行動則で構成された歩行は、ノイズ及び時間遅れのどちらの場合についても歩行動作を保つ事は出来なかったが、図４に示した学習システムでは、ノイズ及び時間遅れがある場合でも歩行が可能であることがわかる。

【０１４４】

よって内部状態を持つ行動則を構成することによって、センサノイズや時間遅れに対してロバストなコントローラを構成出来ることが分かる。

【０１４５】

（３．正規化ガウス関数ネットワークによる関数近似）

２－３－１で述べた価値関数、フィードバック制御器を表現するために用いた、正規化ガウス関数ネットワークについて、以下説明する。

【０１４６】

ＮＧｎｅｔは３層のネットワークで構成されており、中間素子は正規化ガウス関数である。

入力ベクトル$x=(x_1, \cdots, x_n)^T$に対して、ｋ番目のユニットの活性化関数は、以下の式のようになる。

【０１４７】

【数３２】

$$\phi_k(\mathbf{x}) = e^{-\frac{1}{2}\|M_k(\mathbf{x}-\mathbf{c}_k)\|^2} \qquad (42)$$

【０１４８】

ただし、c_kは活性化関数の中心であり、M_kは活性化関数の形状を決定する行列である。ここで、活性化関数ϕ_k（ｘ）を各点で総和が１になるように以下の式（４３）のように正規化したものを、基底関数b_k（ｘ）とする。

【０１４９】

【数３３】

$$b_k(\mathbf{x}) = \frac{\phi_k(\mathbf{x})}{\sum_{l=1}^{K}\phi_l(\mathbf{x})} \qquad (43)$$

【０１５０】

ただし、Ｋは基底関数の個数である。

【０１５１】

このような正規化を行うことによって、中心点c_kが密に配置されている部分では、b_k（ｘ）は局所的な基底関数となり、c_kの分布の端の部分ではb_k（ｘ）はシグモイド関数のような大域的な基底関数になる。

【０１５２】

ネットワークの出力は、基底関数と重みの内積によって以下の式（４４）ようになる。

【０１５３】

【数３４】

$$y_k(\mathbf{x}) = \sum_{k=1}^{K} w_k b_k(\mathbf{x}) \qquad (44)$$

【０１５４】

この出力が、図４の正規化処理部１０３０の出力となる。

【０１５５】

［実施の形態１の変形例］

以上の説明では、図３（３）の構成による制御について説明した。以下では、実施の形態１の変形例として、図３（２）の構成による制御について説明する。

【０１５６】

図１２は、図３（２）の構成に相当するシステムであって、制御器と制御対象を含めたシステム全体の構成を示す図である。

【０１５７】

図１２において、状態観測器２００２は、状態観測器のダイナミクス２００４と、方策勾配法（強化学習）に基づいた強化学習器２００６によって構成される。

【０１５８】

状態観測器２００２中の強化学習器２００６は、以下に説明するとおり、制御対象の観測出力ｙと、状態観測器のダイナミクス２００４に基づく出力と、制御器２０１０の制御出力ｕとに基づいて、学習器出力Ｕを出力する。出力関数処理部２０３０は、状態観測器２００２からの推定状態に基づいて、状態観測器２００２の出力を報酬演算部２０２０に与える。報酬演算部２０２０は、状態観測器２００２の出力と観測対象からの観測出力ｙと学習器出力Ｕとに基づいて、報酬を計算し、強化学習器２００６に与える。

【０１５９】

（方策勾配法を用いた状態観測器の学習）

以下では、実施の形態１の変形例の状態観測器２００２の構造について説明する。

【０１６０】

状態の推定値を、ｘの頭部に“＾”を付加して表現（＝ｘｉ）（以下、本文中では「ｘハット」と呼ぶ）したとき、つぎのような状態観測器を考える。

【０１６１】

【数３５】

$$\dot{\hat{\mathbf{x}}} = \mathbf{f}(\hat{\mathbf{x}}, \mathbf{u}(\hat{\mathbf{x}})) + \mathbf{U}(\hat{\mathbf{x}}, \mathbf{u}(\hat{\mathbf{x}}), \mathbf{y})$$

$$= \mathbf{g}(\hat{\mathbf{x}}, \mathbf{u}(\hat{\mathbf{x}}), \mathbf{U})$$

【０１６２】

ここではまず、通常のオブザーバやカルマンフィルタ同様、対象のダイナミクスｆ（ｘ，ｕ）は既知または学習によって獲得可能であるとし、対象システムの観測出力ｙを基にして、現在の推定状態ｘハットと制御出力ｕから、推定状態を真の状態にどのように近づけるべきかを方策勾配法を用いて学習する。

【０１６３】

ここでは学習器の目的を、状態観測器の出力ｙハット（ｙの頭部に“＾”を付加したもの）と対象システムの出力ｙとの誤差を最小にすることとする。

【０１６４】

よって、報酬演算部２０２０により演算される報酬関数は次のようになる。

【０１６５】

【数３６】

$$r(t) = -(\mathbf{y}(t) - \hat{\mathbf{y}}(t))^T Q(\mathbf{y}(t) - \hat{\mathbf{y}}(t)) - \mathbf{U}^T(t) R \mathbf{U}(t)$$

【０１６６】

ただし、Ｑ，Ｒは報酬関数の形を決めるパラメータである。この結果、学習器は状態観測器のダイナミクス２００４への以下のようなノーテーションのフィードバック入力Ｕを獲得することになる。

【０１６７】

【数３７】

$$\boldsymbol{U}(\hat{\boldsymbol{x}}, \boldsymbol{u}, \boldsymbol{y})$$

【０１６８】

ここで、フィードバック入力Ｕは次の確率分布により表現される。

【０１６９】

【数３８】

$$\pi(U_j, \hat{\mathbf{x}}, \mathbf{y}; \mathbf{w}) = \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left(\frac{-(U_j - \mu_j)^2}{2\sigma_j^2}\right)$$

【０１７０】

したがって、ｊ番目の出力の実現値U_jは、以下の式により与えられる。

【０１７１】

【数３９】

$$U_j(t) = \mu_j(t) + \sigma_j n_j(t)$$

【０１７２】

ただし、n_j（ｔ）～Ｎ（０，１）であり、Ｎ（０，１）は、上述のとおり、平均０、分散１の正規分布を表す。フィードバック入力Ｕを生成する確率分布πの更新は、二足歩行運動の学習の場合と同様に行われる。

【０１７３】

このような構成によっても、周期運動に対し状態観測器の学習を容易とすることが可能な動的制御装置およびこのような動的制御装置を用いた２足歩行移動装置を提供することができる。

【０１７４】

［実施の形態２（多自由時系への適用）］

（５．１）　５リンク２足歩行ロボットモデル

実施の形態２では、図４で構成した学習システムを，多自由度系へ適用した構成について説明する。

【０１７５】

実施の形態１で述べた３リンク２足歩行ロボットモデルに対しての学習結果から、学習手法に方策勾配法を用いることにより、物理系の状態のみを考慮することで、望ましい行動則を獲得出来ることが分かる。

【０１７６】

図１３は、実施の形態２で扱う５リンク２足歩行ロボットおよびそのシミュレーションモデルを説明するための図である。

【０１７７】

実施の形態１の結果を基にして、以下では、図１３（ａ）に示した膝関節を持つ５リンクの実ロボットのシミュレーションモデル（図１３（ｂ））に対して、実施の形態１で提案した手法を適用し、歩行運動の獲得を行なう。

【０１７８】

図１３（ｂ）に示すとおり、実施の形態２のシミュレーションモデルでは、右下腿１０ｒｄ（リンク１）、右上腿１０ｒｕ（リンク２）、腰部４０（リンク２）、左上腿１０ｌｕ（リンク４）および左下腿１０ｌｄ（リンク５）の５リンク系である。

【０１７９】

図１３（ｂ）に示す２足歩行ロボットも、動的制御装置１００と、動的制御装置１００の上部に設けられる胴部４０と、動的制御装置１００により駆動制御される脚部とを備える。脚部は、右上腿１０ｒｕと、右下腿１０ｒｄと、左上腿１０ｌｕと、左下腿１０ｌｄとを有する。

【０１８０】

図１３（ｂ）に示す２足歩行ロボットは、各脚において、右足および左足のそれぞれ接地面近傍に設けられるセンサ２０ｒおよび２０ｌと、右膝部分に設けられるセンサ２０ｒｋと、左膝部分に設けられるセンサ２０ｌｋとを備えるものとする。一方、胴部４０には、センサ（図示せず）が設けられる。動的制御装置１００は、実施の形態１と同様に、駆動部１０８により、右上腿１０ｒｕおよび左上腿１０ｌｕを駆動する。また、左右の膝にも、図示しない駆動部が設けられ、動的制御装置１００からのトルク信号に基づいて、右下腿１０ｒｄおよび左下腿１０ｌｄを駆動するものとする。

【０１８１】

センサ２０ｒｋは、胴部４０の中心線４に対する右上腿１０ｒｕの角度θ^{r}_{hip}、角速度$d\theta^{r}_{hip}/dt$という情報を検出し、また、センサ２０ｌｋは、中心線４に対する左上腿１０ｌｕの角度θ^{l}_{hip}、角速度$d\theta^{l}_{hip}/dt$という情報とを検出し、それぞれ、動的制御装置１００に通知する。さらに、センサ２０ｒは、右上腿１０ｒｕの延長線６ｒに対す

る右脚の角度θ^{r}_{knee}、角速度$d\theta^{r}_{knee}/dt$という情報と右足の接地状態とを検出し、また、センサ２０１は、左上腿１０１ｕの延長線６１に対する左脚の角度θ^{l}_{knee}、角速度$d\theta^{l}_{knee}/dt$という情報と左足の接地状態とを検出し、それぞれ、動的制御装置１００に通知する。

【０１８２】

また、胴部４０に設けられるセンサは、鉛直方向２に対する胴部４０のピッチ角θ_{p}、角速度$d\theta_{p}/dt$という情報を検出し、それぞれ、動的制御装置１００に通知する。

【０１８３】

５リンク２足歩行ロボットモデルの物理パラメータを表１に示す。

【０１８４】

【表２】

	リンク1	リンク2	リンク3	リンク4	リンク5
質量 [kg]	0.15	0.64	2.0	0.64	0.15
長さ [m]	0.20	0.20	0.01	0.20	0.20
慣性モーメント $(\times 10^{-4})[kg \cdot m^2]$	1.40	6.89	1.0	6.89	1.40
各関節から重心までの長さ [m]	0.1	0.12	0.0	0.12	0.1

【０１８５】

（５．２） 多自由度系に対する学習システム

図１３（ｂ）に示す５リンク２足歩行ロボットモデルに対する学習システムについて、以下さらに詳しく説明する。

【０１８６】

動的制御装置１００では、膝関節の状態は学習に用いず、実施の形態１の図４と同様に、腰関節の状態（θ^{r}_{hip}、θ^{l}_{hip}、$d\theta^{r}_{hip}/dt$、$d\theta^{l}_{hip}/dt$）及びピッチ角の状態（θ_{p}、$d\theta_{p}/dt$）のみに関する状態変数を用いて学習を行う。

【０１８７】

一方、膝関節については、接地情報と腰関節の状態を基に膝関節の目標関節角を切替える状態マシン１０４０を制御器として用いる。この状態マシン１０４０については後述する。

【０１８８】

図１４は、動的制御装置１００における実施の形態２の学習システムの構成を示す図である。なお、図４に示した実施の形態１の動的制御装置１００と同一部分は、同一符号で示す。ただし、図１４においては、フィードバック制御器１０２２´は、フィードバック制御器１０２２の機能と出力飽和処理部１０２４の機能とを併せて有するものとし、価値関数処理部１０３２´は、価値関数処理部１０３２の機能と正規化処理部１０３０の機能も併せて有するものとする。

【０１８９】

左上腿１０１ｕおよび右上腿１０ｒｕへのトルクは、ＰＤサーボ処理部１０２８から与えられる。

【０１９０】

したがって、価値関数処理部１０３２´、フィードバック制御器１０２２´、ＣＰＧ処理部１０２６およびＰＤサーボ処理部１０２８の動作は、駆動する対象が異なるのみで、その基本的な動作は、実施の形態１と同様である。

【０１９１】

一方、膝関節へのトルク入力は、状態マシン１０４０によって決定される目標関節角を用いたＰＤサーボ処理部１０４２により与えられる。

【０１９２】
（状態マシン１０４０による膝関節制御器）
図１５は、状態マシン１０４０の動作を説明するための概念であり、図１５（ａ）は、右膝の状態を、図１５（ｂ）は左膝の状態をそれぞれ示す。
【０１９３】
状態は、膝屈曲→膝伸長→膝屈曲→膝伸長との状態遷移を繰り返す。
【０１９４】
たとえば、右膝では、前半の膝屈曲→膝伸長は、脚の「振り状態」であり、目標角度はそれぞれθ_1およびθ_2である。右膝の後半の膝屈曲→膝伸長は、脚の「立ち状態」であり、目標角度はそれぞれθ_3およびθ_4である。
【０１９５】
一方で、左膝では、前半の膝屈曲→膝伸長は、脚の「立ち状態」であり、目標角度はそれぞれθ_3およびθ_4である。左膝の後半の膝屈曲→膝伸長は、脚の「振り状態」であり、目標角度はそれぞれθ_1およびθ_2である。
【０１９６】
腰関節の角度θ^r_{hip}、θ^r_{hip}（以下、総称するときは、「θ_{hip}」）および接地情報を用いて、図１５に示す状態マシンにより，膝関節の目標角度（θ^r_{knee}ハット、θ^l_{knee}ハット：総称するときは、「θ_{knee}ハット」）を決定し、以下に示すＰＤサーボにより膝関節へのトルクu_{knee}を出力する。
【０１９７】
【数４０】

$$u_{knee} = K_p(\hat{\theta}_{knee} - \theta_{knee}) - K_d\dot{\theta}_{knee}$$

【０１９８】
ここで、K_pは位置ゲイン、K_dは速度ゲインである。ただし、シミュレーションでは、K_p＝１２．０、K_d＝０．１５とした。
【０１９９】
図１５に示すように、状態マシンでは４つの目標角（θ_1、θ_2、θ_3、θ_4）を設定した条件に応じて切り換える。ここでは、θ_1＝１．１１，θ_2＝０．５６，θ_3＝０．５２，θ_4＝０．２６とした。膝を曲げた状態から伸ばした状態への遷移は腰関節を用いた以下の条件に基づいて行なわれる。
【０２００】
【数４１】

$$\theta^l_{hip} - \theta^r_{hip} < b \quad \text{または} \quad \theta^r_{hip} - \theta^l_{hip} < b$$

【０２０１】
膝を伸ばした状態から曲げた状態への遷移は足裏の接地情報により行われる。シミュレーションでは、たとえば、ｂ＝０．１５とする。
【０２０２】
（床反力モデル）
以下では、さらに、シミュレーションで用いる床反力モデルについて説明する。
【０２０３】
ｘ、ｙは脚端の位置を表し、ｘｇ、ｙｇは接地点とすると、接地時の床反力は以下の式でモデル化される。
【０２０４】

【数４２】

$$F_x = k_x(x_g - x) - b_x\dot{x}$$

$$F_y = k_y(y_g - y) - b_y\dot{y}$$

$$if\ y \geq y_g$$

$$F_x = 0, F_y = 0$$

【０２０５】

ここで、F_x、F_yは水平方向、垂直方向の床反力である。

【０２０６】

以下で説明するシミュレーションでは、上記の式のそれぞれの係数はk_x＝３０００、b_x＝１０、k_y＝３００００、b_y＝１００とする。また、床反力が$F_x > \mu F_y$を満たすときに、足裏が床面を滑べると定義し、ここでのμは静摩擦係数であり、μ＝１．０としている。

【０２０７】

（５．３）シミュレーション

以下、実施の形態２でのシミュレーション結果を説明する。

【０２０８】

本シミュレーションで用いるＣＰＧのパラメータは、ロボットモデルの脚長の変化を考慮して、図４の３リンク２足歩行ロボットモデルで用いた特性と比較して、低い周波数が必要となるので，各パラメータを以下のようにする。

【０２０９】

τ＝０．４、τ′＝０．８、β＝１．３、ω＝２．０、z_0＝０．１

正規化ガウス関数ネットワークＮＧｎｅｔの基底関数については、それぞれの状態変数に対して格子状に、ｘ＝（x_1，x_2，x_3，x_4）^{T}を以下のように配置する。

【０２１０】

【数４３】

$$x = (x_1, x_2, x_3, x_4)^T$$

$$= (\theta_{hip}^l + \theta_p, \theta_{hip}^r + \theta_p, \dot{\theta}_{hip}^l + \dot{\theta}_p, \dot{\theta}_{hip}^r + \dot{\theta}_p)^T : (16, 10, 16, 10)$$

【０２１１】

基底関数は、２５６００個をそれぞれの以下の範囲に均等に配置した。

【０２１２】

【数４４】

$$(-\frac{\pi}{3} \leq x_1 \leq \frac{\pi}{3}, -2.5\pi \leq x_2 \leq 2.5\pi, -\frac{\pi}{3} \leq x_3 \leq \frac{\pi}{3}, -2.5\pi \leq x_4 \leq 2.5\pi)$$

【０２１３】

報酬関数に関しては、式（４１）を使用するが、物理パラメータの変更にともない、ｋｓ＝０．０６、k_H＝０．５、ｈ′＝０．３とした。

【０２１４】

その他の学習パラメータ等については、式（３８）のβσ（σは上付き）＝０．０２以外は、実施の形態１と全て同様とする。

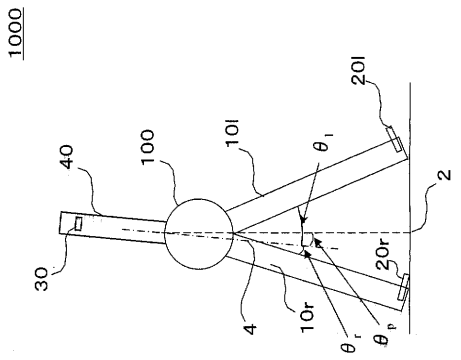
【０２１５】

また、１学習試行の終了条件についても、実施の形態１と同様以下の条件を用いる。

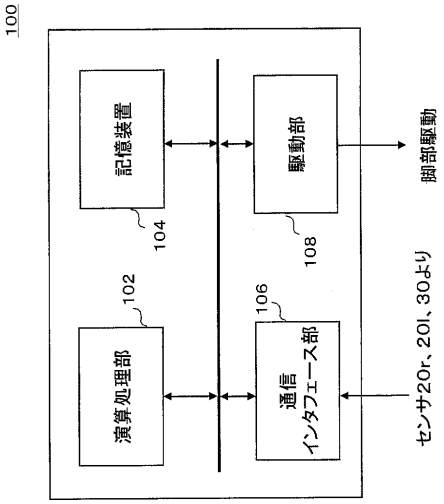
【０２１６】
・１７７００ｍｓｅｃ経過
・転倒時（ただし、同時にr＝－1の報酬を与える）
（５．４） 学習結果
実施の形態２のシミュレーションにおいても、３リンク２足歩行ロボットでの学習と同様、ロボットが１０試行連続で、転倒せずに歩き続けたとき、２足歩行運動を獲得したと定義する。１０回シミュレーションを行った結果全ての試行で2足歩行運動の獲得に成功した。また、運動獲得に必要な平均試行回数は８９９回であった。
【０２１７】
図１６は、学習過程の一例の試行回数と獲得報酬の総和の関係を示す図であり、図１７は、学習前、学習後の歩行軌跡を示す図である。
【０２１８】
図１７において、（１）は学習前の歩行軌跡であり、（２）は１５００回学習後の歩行軌跡を示している。
【０２１９】
この結果は、２足歩行ロボットの自由度が増加し、状態空間が高次元となった場合でも、少ない状態変数だけを観測することで，少ない試行回数で目的の周期運動を獲得できることがわかる。
【０２２０】
図１６の結果からすると、５リンクの２足歩行ロボットに対して、強化学習を用いた従来例（たとえば、文献：Y.Nakayama, M.Sato, and S. Ishii. Reinforcement Learning for biped robot. In Proceedings of the 2nd International Symposium on Adaptive Motion of Animals and Machines, pp. Thp-II-5, 2003.）と比較しても、かなり少ない試行回数で歩行運動を獲得できている。
【０２２１】
以上のような構成によって、２足歩行ロボットの自由度が増加し、状態空間が高次元となった場合でも、周期運動に対し状態観測器の学習を容易とすることが可能な動的制御装置およびこのような動的制御装置を用いた２足歩行移動装置を提供することができる。
【０２２２】
今回開示された実施の形態はすべての点で例示であって制限的なものではないと考えられるべきである。本発明の範囲は上記した説明ではなくて特許請求の範囲によって示され、特許請求の範囲と均等の意味および範囲内でのすべての変更が含まれることが意図される。
【図面の簡単な説明】
【０２２３】
【図１】本発明の動的制御装置を用いた２足歩行移動システム１０００の一例を示す概念図である。
【図２】図１に示した動的制御装置１００の構成を示すブロック図である。
【図３】動的行動則を説明するための概念図である。
【図４】演算処理部１０２の行う処理を示す機能ブロック図である。
【図５】神経振動子モデルによるＣＰＧを示す概念図である。
【図６】ＣＰＧの出力ｑを構成する変数z_1、z_2の波形を示す図である。
【図７】１試行で獲得した報酬の総和を、試行回数ごとに取った学習曲線を示す図である。
【図８】図７の学習曲線に対応する歩行の軌跡を示す図である。
【図９】ＣＰＧの内部状態とロボットの状態の間のリミットサイクルを、ＣＰＧの内部状態z1と脚角度の時間変化として示す図である。
【図１０】ＣＰＧの内部状態とロボットの状態の間のリミットサイクルを、脚角度、脚の角速度、ＣＰＧの内部状態z_1の関係として示す図である。
【図１１】センサノイズ・時間遅れに対するシミュレーション結果を示す図である。

【図１２】制御器と制御対象を含めたシステム全体の構成を示す図である。
【図１３】実施の形態２で扱う５リンク２足歩行ロボットおよびそのシミュレーションモデルを説明するための図である。
【図１４】動的制御装置１００における実施の形態２の学習システムの構成を示す図である。
【図１５】状態マシン１０４０の動作を説明するための概念である。
【図１６】学習過程の一例の試行回数と獲得報酬の総和の関係を示す図である。
【図１７】学習前、学習後の歩行軌跡を示す図である。
【符号の説明】
【０２２４】
１０ｒ，１０ｌ　脚部、２０ｒ，２０ｌ，３０　センサ、４０　胴部、１００　動的制御装置、１０２　演算処理部、１０４　記憶装置、１０６　通信インタフェース、１０８　駆動部、１０００　２足歩行移動システム。

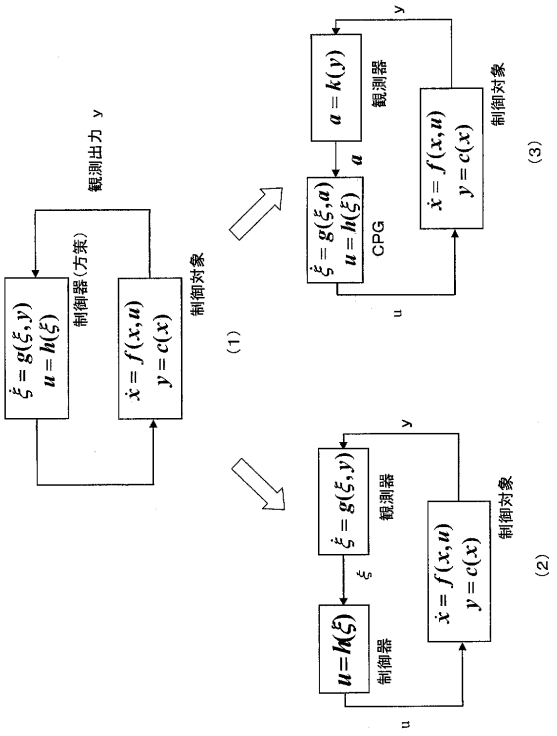
【図１】



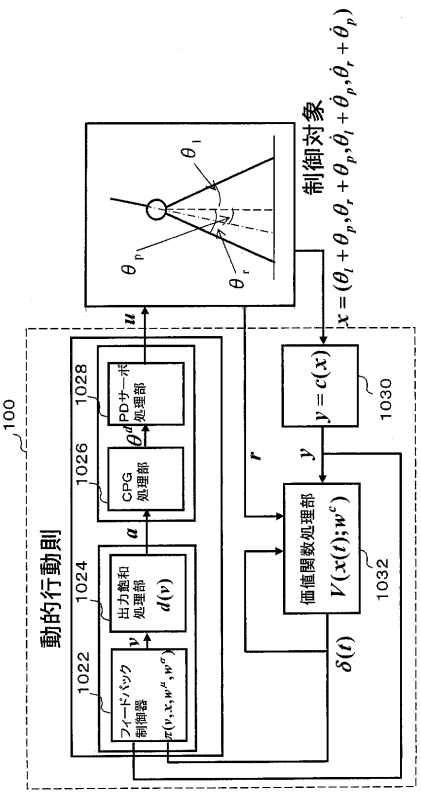
【図２】



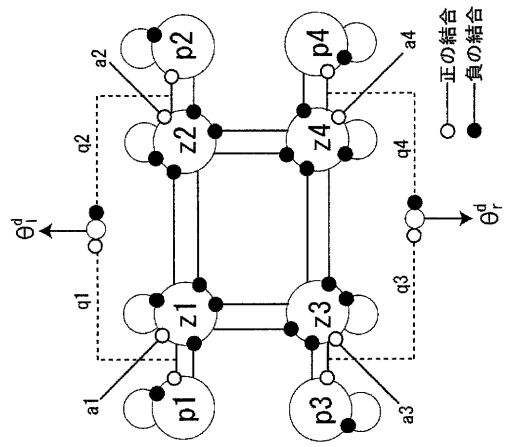
【図３】


観測出力 y
制御器（方策）
制御対象
(1)
CPG
観測器
制御器
制御対象
(2)
(3)

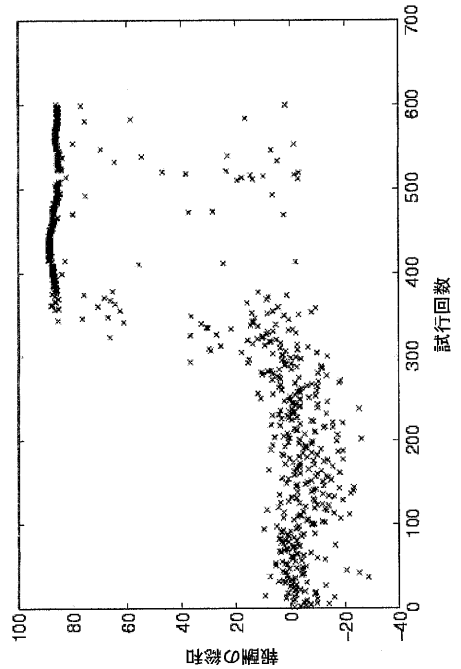
【図４】


100
動的行動則
1022
1024
1026
1028
1030
1032
制御対象

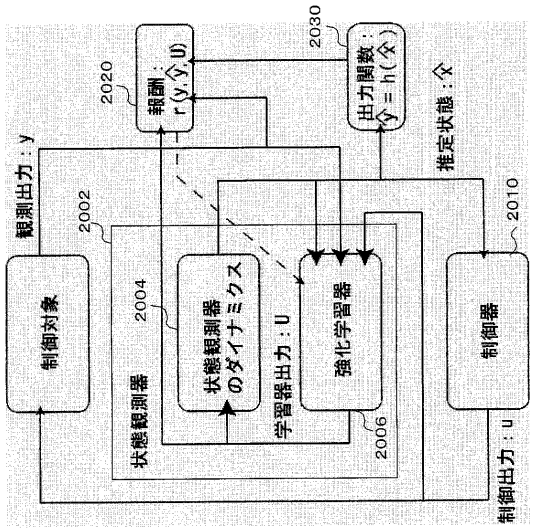
【図５】


正の結合
負の結合
p1
p2
p3
p4
z1
z2
z3
z4
a1
a2
a3
a4
q1
q2
q3
q4

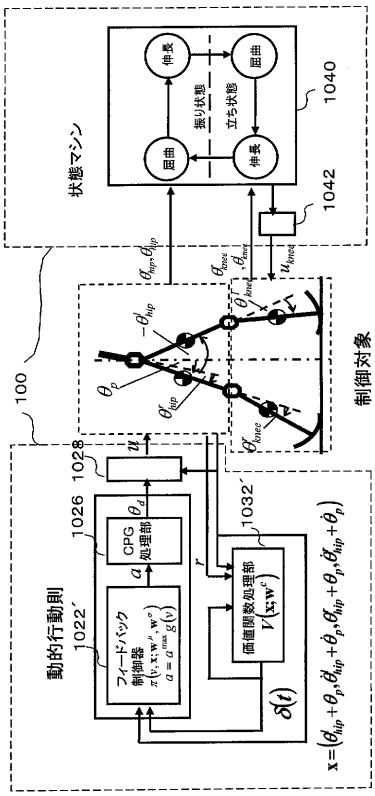
【図７】


報酬の総和
試行回数

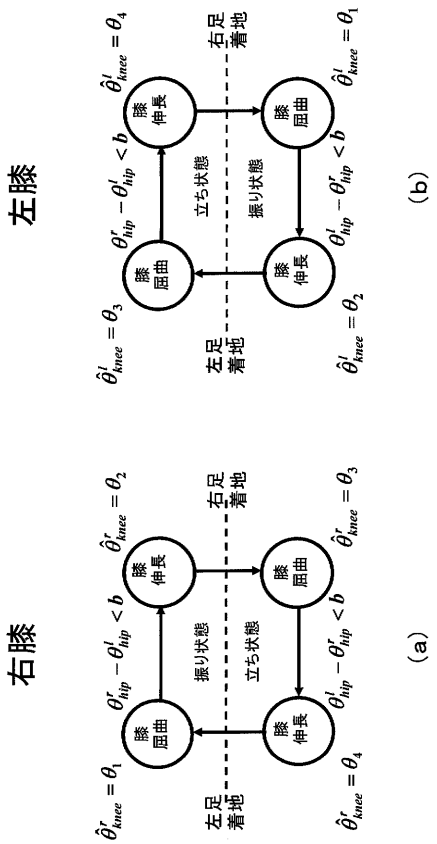
【図１２】



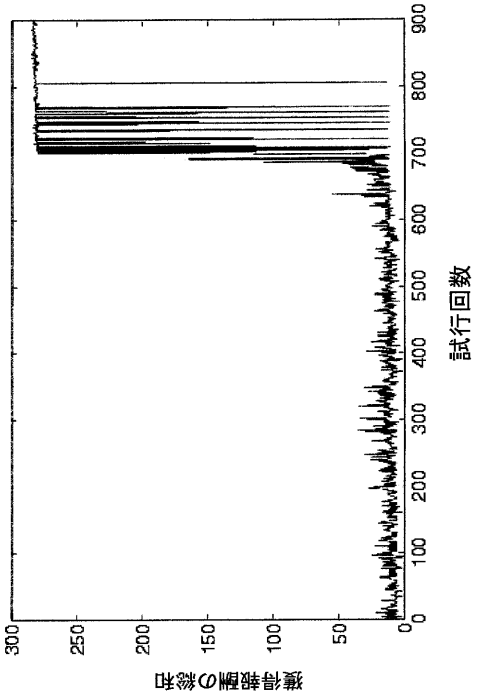
【図１４】



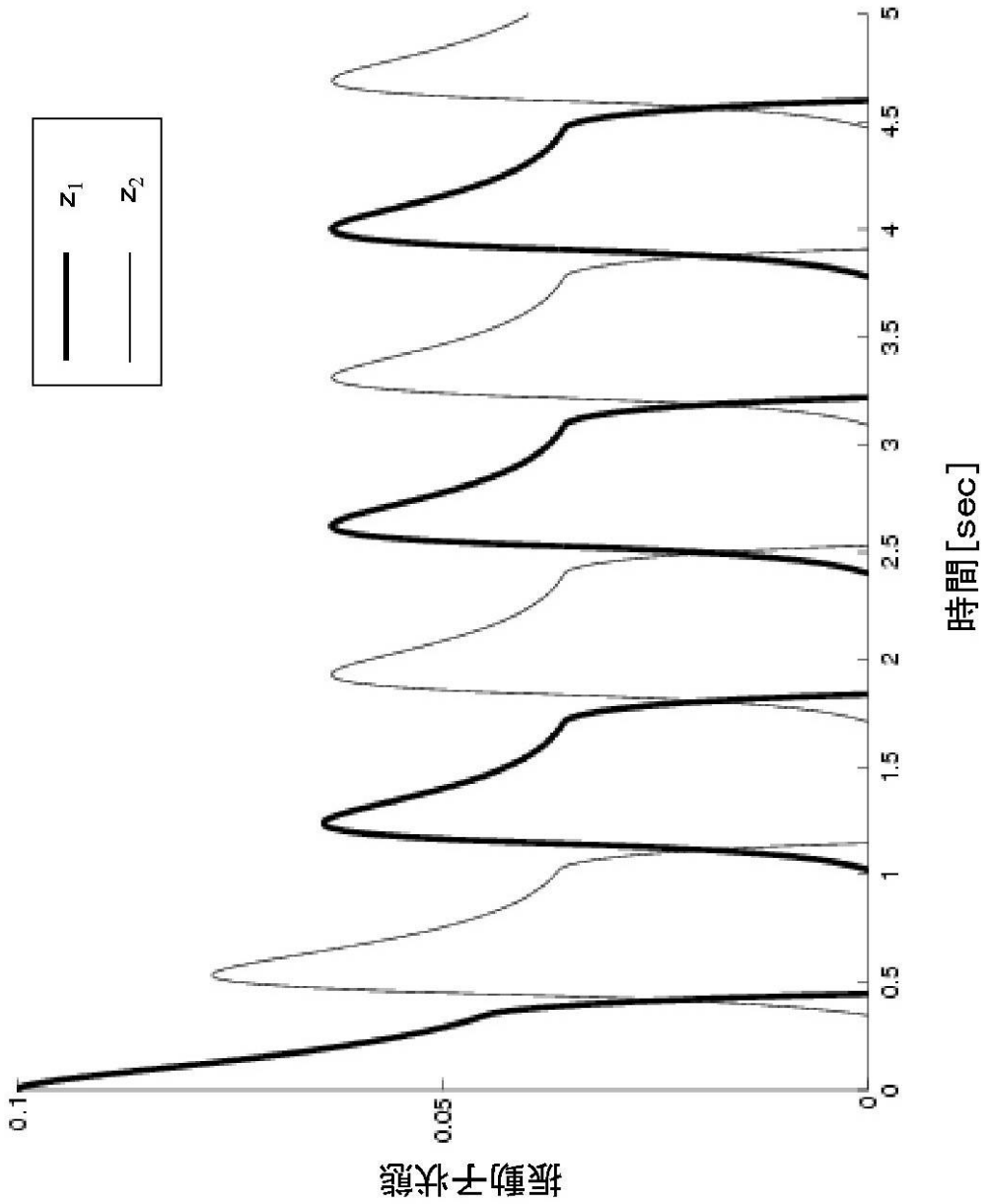
【図１５】



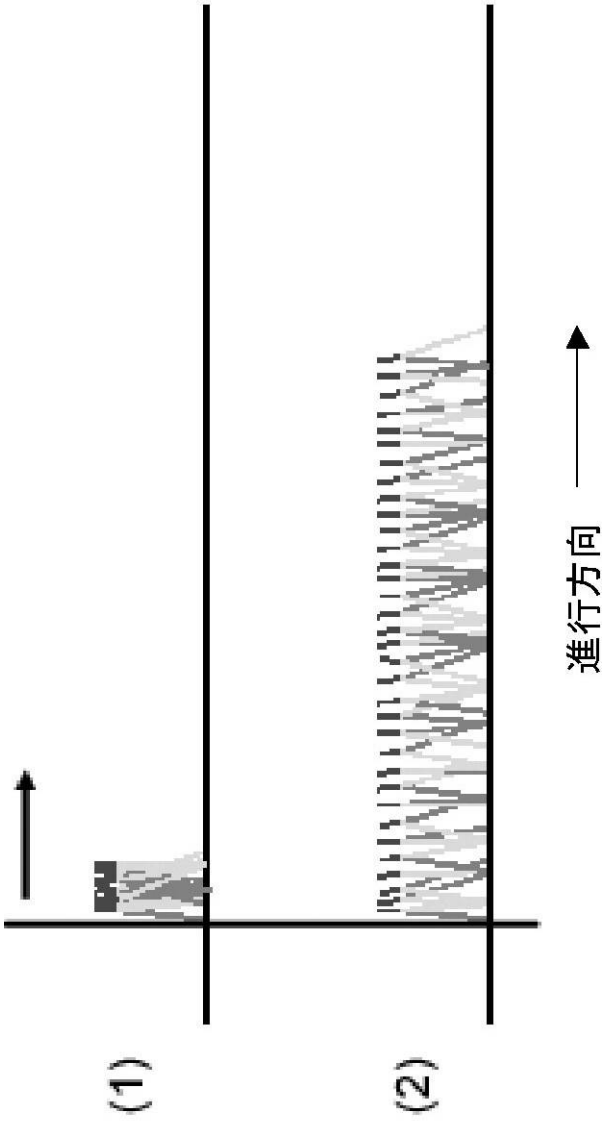
【図１６】



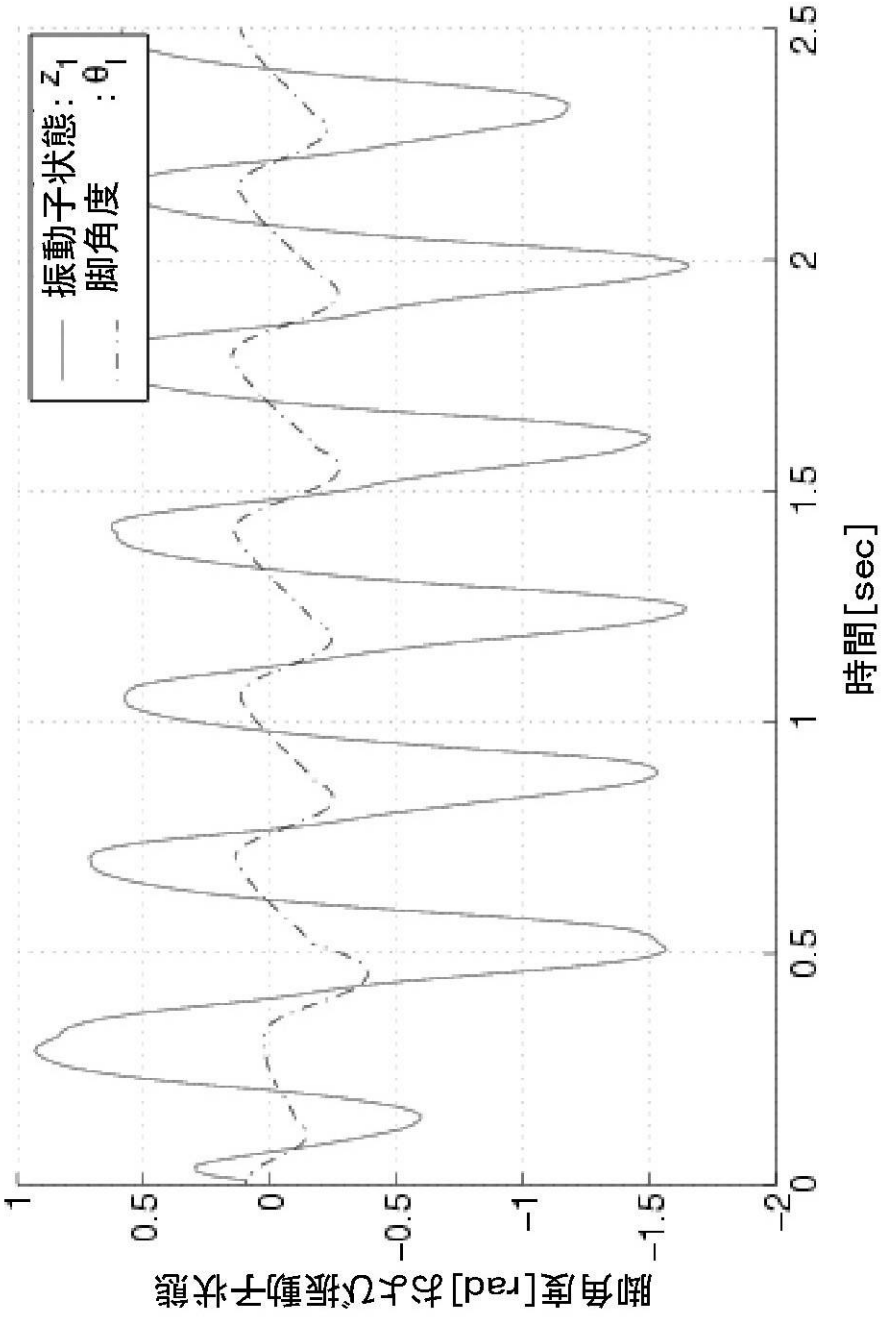
【図６】



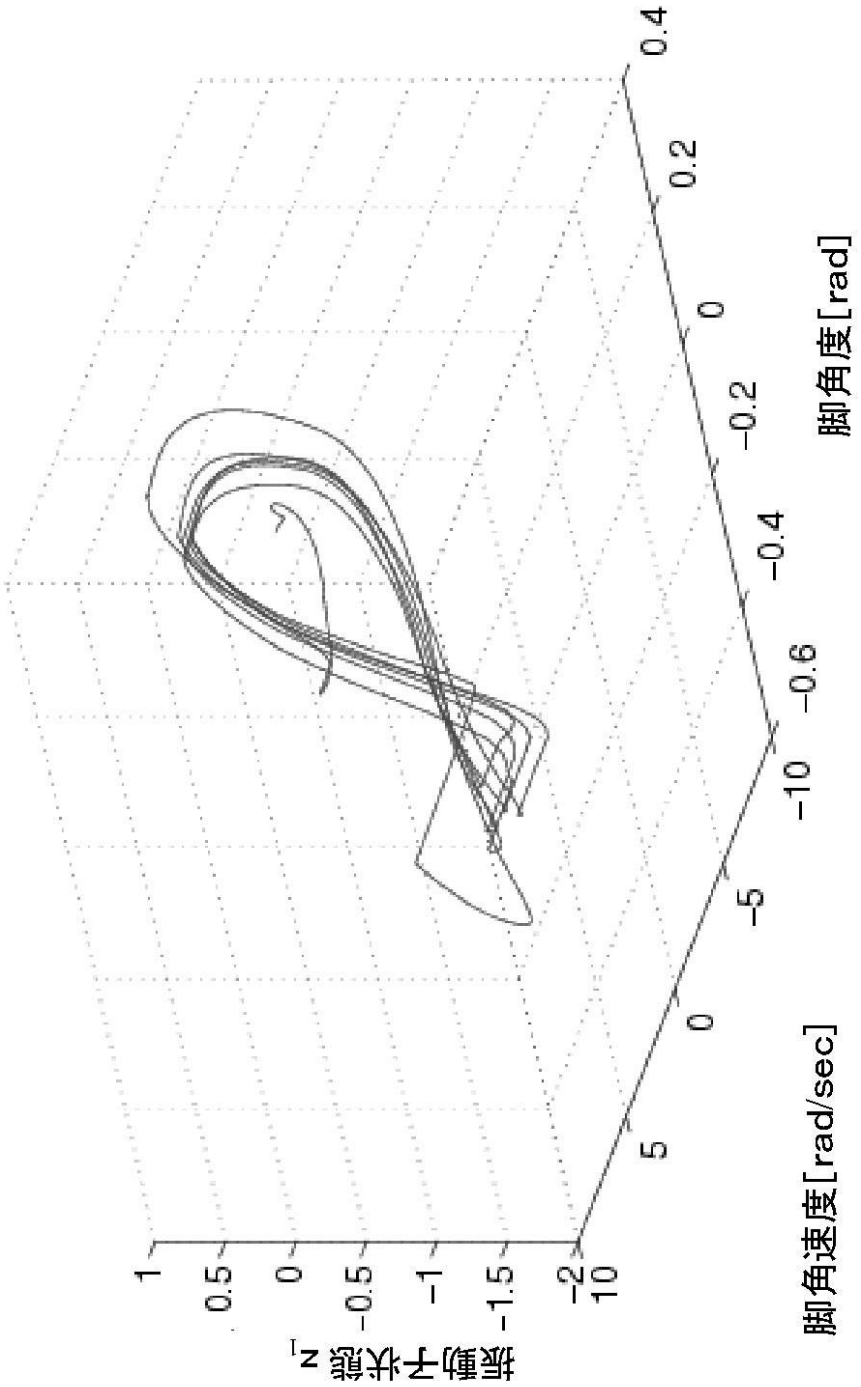
【図８】



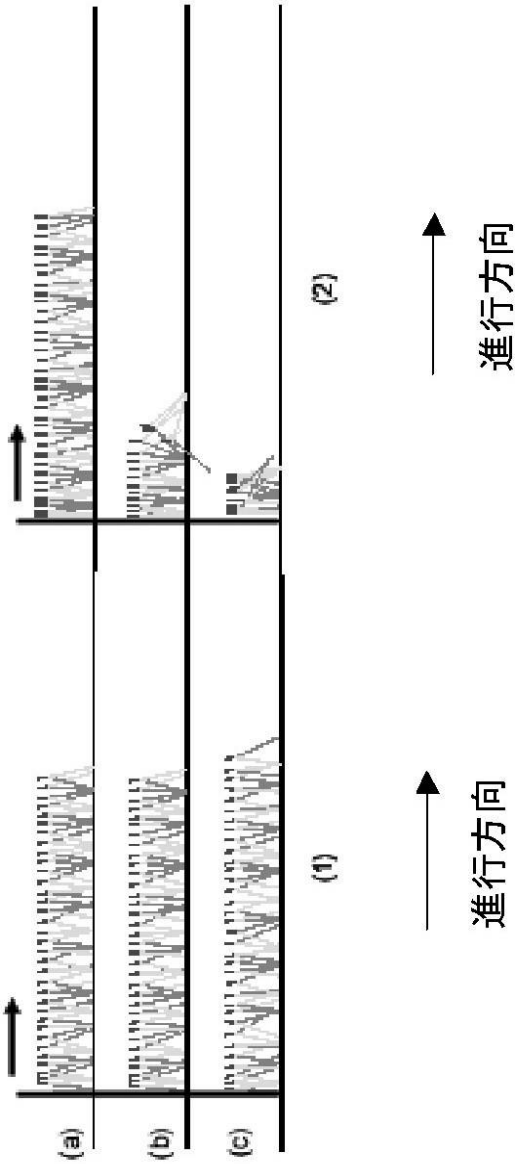
【図９】


振動子状態：z_1
脚角度　：θ_1
脚角度[rad]および振動子状態
時間[sec]
1
0.5
0
−0.5
−1
−1.5
−2
0
0.5
1
1.5
2
2.5

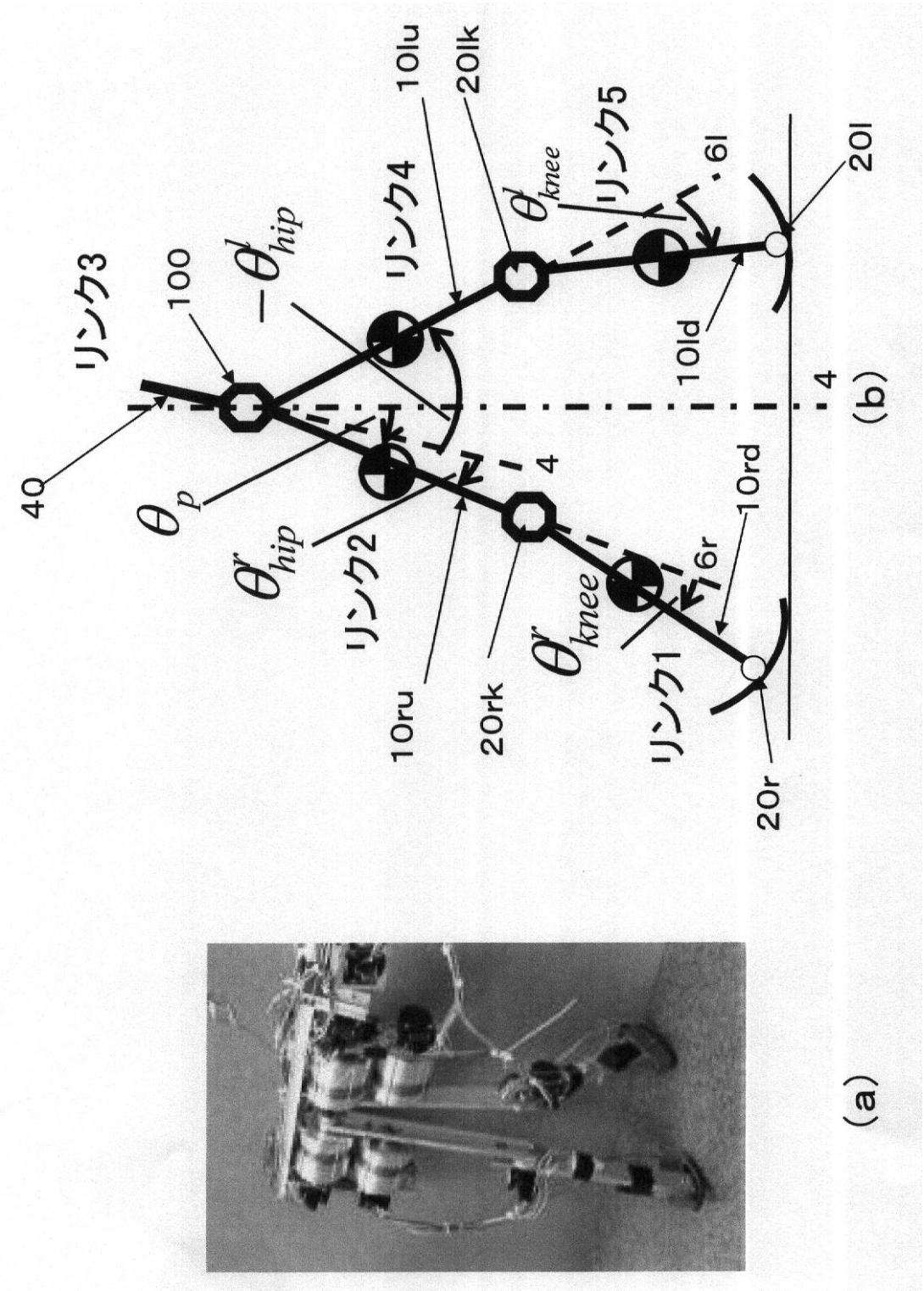
【図１０】


振動子状態 z_1
1
0.5
0
−0.5
−1
−1.5
−2
10
5
0
−5
−10
脚角速度[rad/sec]
0.4
0.2
0
−0.2
−0.4
−0.6
脚角度[rad]

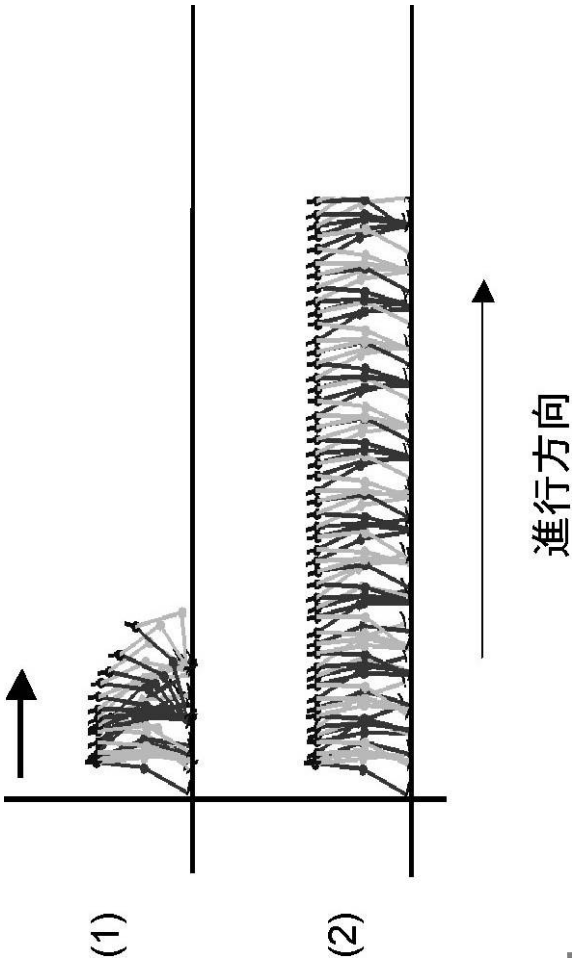
【図１１】



【図１３】


リンク3
100
$-\theta^l_{hip}$
リンク4
101u
201k
θ^l_{knee}
リンク5
61
201
101d
4
(b)
4Q
θ_p
θ^r_{hip}
リンク2
4
101u
201k
θ^r_{knee}
リンク1
61
101d
201r
(a)

【図１７】



フロントページの続き

(74)代理人 100109162
弁理士 酒井 將行
(72)発明者 森本 淳
京都府相楽郡精華町光台二丁目２番地２ 株式会社国際電気通信基礎技術研究所内
(72)発明者 松原 崇充
京都府相楽郡精華町光台二丁目２番地２ 株式会社国際電気通信基礎技術研究所内
(72)発明者 佐藤 雅昭
京都府相楽郡精華町光台二丁目２番地２ 株式会社国際電気通信基礎技術研究所内
(72)発明者 中西 淳
京都府相楽郡精華町光台二丁目２番地２ 株式会社国際電気通信基礎技術研究所内
(72)発明者 遠藤 玄
東京都品川区北品川６丁目７番３５号 ソニー株式会社内

審査官 植村 森平

(56)参考文献 森健、中村泰、石井信，二足歩行運動に対する方策勾配法に基づいた強化学習法，電子情報通信学会技術研究報告，日本，社団法人 電子情報通信学会，２００４年 3月１９日，Vol．103，No．734，pp．73-78
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Ｄ－２，日本，社団法人 電子情報通信学会，２００４年 3月 1日，Vol．J87-D-2，No．3，pp.893-902
松原崇充、森本淳、中西淳、佐藤雅昭、銅谷賢治，方策勾配法を用いた動的行動則の獲得：２足歩行運動への適用，電子情報通信学会技術研究報告，日本，社団法人 電子情報通信学会，２００４年 1月２７日，Vol．103，No．602，pp．53-58
Marc H．Raibert，LEGGED ROBOTS，Communications of the ACM，米国，Association for Computing Machinery，１９８６年 6月，Vol．29，issue 6，pp．499-514
中村泰、石井信、佐藤雅昭，神経振動子ネットワークを用いた強化学習法による歩行運動の獲得，電気情報通信学会技術研究報告，日本，社団法人 電子情報通信学会技術研究報告，２００２年 3月１８日，Vol．101，No．735，pp．183-190
森健、吉本潤一郎、石井信，確率的方策勾配法に基づくactor-critic法と連続システムの制御への応用，電子情報通信学会技術研究報告，日本，社団法人 電子情報通信学会，２００３年 3月１９日，Vol．102，No．731，pp．137-142

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｂ２５Ｊ ５／００
ＪＳＴＰｌｕｓ（ＪＤｒｅａｍＩＩ）
ＣｉＮｉｉ