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

特許第4630553号
(P4630553)

(45)発行日　平成23年2月9日(2011.2.9)　(24)登録日　平成22年11月19日(2010.11.19)

(51)Int.Cl.		F I		
B25J 5/00	(2006.01)	B25J	5/00	F
B25J 13/00	(2006.01)	B25J	13/00	Z
G06N 3/00	(2006.01)	G06N	3/00	55OE

請求項の数 5　(全 22 頁)

(21)出願番号　特願2004-8130(P2004-8130)
(22)出願日　平成16年1月15日(2004.1.15)
(65)公開番号　特開2005-199383(P2005-199383A)
(43)公開日　平成17年7月28日(2005.7.28)
審査請求日　平成19年1月12日(2007.1.12)

(73)特許権者　000002185
ソニー株式会社
東京都港区港南１丁目７番１号
(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　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

最終頁に続く

(54)【発明の名称】動的制御装置および動的制御装置を用いた２足歩行移動体

(57)【特許請求の範囲】
【請求項１】
制御対象に対する制御信号を生成するための動的制御装置であって、
前記制御対象の状態情報を検知するためのセンサ群と、
周期的な時間発展を行なう内部状態を有し、前記制御対象の状態推定を行なって、前記制御対象に対する前記制御信号を生成する制御手段を備え、
前記制御手段は、
前記センサ群から得られる前記状態情報のみで構成される状態空間を用いた方策勾配法による強化学習を行い、前記制御対象から得られる前記状態情報と前記制御信号により規定される報酬に基づく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し、更新された前記フィードバックパラメータにより規定される確率分布により出力値を決定するフィードバック制御器と、
前記周期的な時間発展を行なう前記内部状態を有し、前記フィードバック制御器からの前記出力値に応じて変化する前記内部状態に基づき、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、前記制御信号を生成するためのセントラルパターンジェネレータとを含み、
前記制御信号に基づいて、前記制御対象を駆動するための駆動手段とを備える、動的制御装置。
【請求項２】
前記確率分布は、前記フィードバックパラメータにより決定される平均値と分散とを有

する正規分布により表され、
前記価値関数は、正規化ガウスネットワークで表現される、請求項１記載の動的制御装置。
【請求項３】
前記平均値は、前記フィードバックパラメータを重みとする正規化ガウス関数ネットワークで表現され、
前記分散は、前記フィードバックパラメータによるシグモイド関数として表現され、
前記フィードバックパラメータは、各前記フィードバックパラメータに対応する学習率とテンポラル・ディファレンス誤差とエリジビリティ・トレースとに応じて更新される、請求項２記載の動的制御装置。
【請求項４】
前記フィードバック制御器は、前記出力値を所定のレベル以下に制限するための出力飽和手段を含む、請求項２記載の動的制御装置。
【請求項５】
請求項１～４のいずれか１項に記載される動的制御装置により、２足歩行制御が行なわれる、２足歩行移動体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、非線形の制御対象に関し、２足歩行等の周期運動や、状態観測器を実現する上で有効となる動的制御装置と動的制御装置を用いた２足歩行移動体とに関する。
【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なう)（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およびこのような動的制御装置を用いた２足歩行移動体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、タスクを行な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なっている。
【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、制御対象に対する制御信号を生成するための動的制御装置であって、制御対象の状態情報を検知するためのセンサ群と、周期的な時間発展を行なう内部状態を有し、制御対象の状態推定を行なって、制御対象に対する制御信号を生成する制御手段を備え、制御手段は、センサ群から得られる状態情報のみで構成される状態空間を用いた方策勾配法による強化学習を行い、制御対象から得られる状態情報と制御信号により規定される報酬に基づく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し、更新されたフィードバックパラメータにより規定される確率分布により出力値を決定するフィードバック制御器と、周期的な時間発展を行なう内部状態を有し、フィードバック制御器からの出力値に応じて変化する内部状態に基づき、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、制御信号を生成するためのセントラルパターンジェネレータとを含み、制御信号に基づいて、制御対象を駆動するための駆動手段とを備える。
【００１０】
好ましくは、確率分布は、フィードバックパラメータにより決定される平均値と分散とを有する正規分布により表され、価値関数は、正規化ガウスネットワークで表現される。
【００１１】
好ましくは、平均値は、フィードバックパラメータを重みとする正規化ガウス関数ネットワークで表現され、分散は、フィードバックパラメータによるシグモイド関数として表現され、フィードバックパラメータは、各フィードバックパラメータに対応する学習率とテンポラル・ディファレンス誤差とエリジビリティ・トレースとに応じて更新される。
【００１２】
好ましくは、フィードバック制御器は、出力値を所定のレベル以下に制限するための出力飽和手段を含む。
【００１３】
好ましくは、この発明の２足歩行移動体は、上記動的制御装置により、２足歩行制御が行なわれる。
【発明の効果】

【００１４】
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わ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なう場合に、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。
【発明を実施するための最良の形態】
【００１５】
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。
【００１６】
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な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、周期運動を行なう劣駆動（機械）系に適用するのに適した制御システムを提供する。
【００１７】
［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。
【００１８】
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。
【００１９】
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。
【００２０】
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。
【００２１】
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。
【００２２】
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の駆動制御を行なうための駆動部１０８とを備える。
【００２３】
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。
【００２４】
（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。
【００２５】
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。

【００２６】
「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わされるダイナミクスを持つ枠組である。
【００２７】
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。
【００２８】
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。
【００２９】
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。
【００３０】
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な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。
【００３１】
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。
【００３２】
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。
【００３３】
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なう。
【００３４】
（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。
【００３５】
図４は、図２で説明した演算処理部１０２の行な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。
【００３６】
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、以下の式で表わされる。
【００３７】
【数１】

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

【００３８】
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、１０１の角度であり、θ_pは胴体４０のピッチ角である。
【００３９】

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。

【００４０】

また、ここでは全状態観測を仮定し、ｙ＝ｘであるとする。

【００４１】

（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１０２２の構成として、以下の式で表わ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.”、Biologial Cybernetics, Vol.52, pp. 367-376, 1985に開示がある。

【００４２】

【数２】

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

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

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

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

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

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

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

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

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

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

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

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

【００４３】

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。

【００４４】

図５は、式（１）～（１２）で表わ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なうものは白丸で、負の結合を行なうものは黒丸で示している。

【００４５】

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。

【００４６】

式（１）～（１２）で表わ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。

【００４７】

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。

【００４８】

【数３】

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

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

【００４９】

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。

【００５０】

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。

【００５１】

【数４】

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

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

【００５２】

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。

【００５３】

（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）によって表わされる。

【００５４】

【数５】

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

【００５５】

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。

【００５６】

【数６】

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

【００５７】

ただし、n_j（ｔ）～Ｎ（０，１）であり、Ｎ（０，１）は平均０、分散１の正規分布を表わす。

【００５８】

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。

【００５９】

【数７】

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

【００６０】

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。

【００６１】

【数８】

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

【００６２】

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。

【００６３】

（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ない、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。

【００６４】

（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）で表わす。

【００６５】

【数９】

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

【００６６】

ただし、ｘ∈Ｘ⊂Ｒ^{n}は状態、ｕ∈Ｕ⊂Ｒ^{m}は制御入力を表わす。

【００６７】

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。

【００６８】

【数１０】

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

【００６９】

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。

【００７０】

【数１１】

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

【００７１】

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。

【００７２】

【数１２】

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

【００７３】

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。

【００７４】

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なう。

【００７５】

【数１３】

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

【００７６】

上式は連続時間系でのＴＤ誤差である。

【００７７】

（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なう場合では、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。

【００７８】

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。

【００７９】

【数１４】

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

【００８０】

ただし、以下の式（２６）が成り立つ。

【００８１】

【数１５】

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

【００８２】

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。

【００８３】

よって、パラメータの更新則は次の式（２７）のようになる。

【００８４】

【数１６】

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

【００８５】

ただし、ηは学習率である。

【００８６】

（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なう。

【００８７】

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なうことで望みの動的行動則を獲得することを考える。

【００８８】

（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。

【００８９】

【数１７】

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

【００９０】

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。

【００９１】

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。

【００９２】

【数１８】

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

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

【００９３】

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。

【００９４】

（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。

【００９５】

【数１９】

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

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

【００９６】

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わし、標準偏差σをシグモイド関数によって表わす。

【００９７】

【数２０】

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

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

【００９８】

ただし、ノーテーションとしては以下のとおりである。

【００９９】

【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

【０１００】

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。

【０１０１】

【数２２】

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

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

【０１０２】

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。

【０１０３】

【数２３】

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

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

【０１０４】

ただし、ノーテーションとしては以下のとおりである。

【０１０５】

【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

【０１０６】

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。

【０１０７】

【数２５】

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

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

【０１０８】

ただし、ノーテーションとしては以下のとおりである。

【０１０９】

【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

【０１１０】

（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なった結果について以下説明する。

【０１１１】

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。

【０１１２】

それぞれの学習パラメータは、以下のとおりである。

【０１１３】

【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 ： $K_p = 5.0$

$K_d = 0.1$

【０１１４】

また、ＮＧｎｅｔの基底関数は、実際にロボットが歩行運動を行な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。

【０１１５】

【数２８】

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

【０１１６】

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。

【０１１７】

【数２９】

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

【０１１８】

報酬関数は以下の式で表わす。

【０１１９】

【数３０】

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

【０１２０】

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、以下の式で表わされる。

【０１２１】

【数３１】

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

【０１２２】

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。

【０１２３】

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。

【０１２４】

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。

【０１２５】

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。

【０１２６】

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。

【０１２７】

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。

【０１２８】

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。

【０１２９】

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。

【０１３０】

また、６００回学習試行を行な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なうことによって、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なうことのよって、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。

【０１３１】

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。

【０１３２】

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。

【０１３３】

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。

【０１３４】

（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。

【０１３５】

【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

【０１３６】
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。
【０１３７】
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。
【０１３８】
（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回学習試行を行なうことによって、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なった。
【０１３９】
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なった。
【０１４０】
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わしている。
【０１４１】
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。
【０１４２】
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。
【０１４３】
（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。
【０１４４】
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。
【０１４５】

【数３２】

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

【０１４６】

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。

【０１４７】

【数３３】

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

【０１４８】

ただし、Ｋは基底関数の個数である。

【０１４９】

このような正規化を行なうことによって、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。

【０１５０】

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。

【０１５１】

【数３４】

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

【０１５２】

この出力が、図４の正規化処理部１０３０の出力となる。

【０１５３】

［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。

【０１５４】

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。

【０１５５】

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。

【０１５６】

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。

【０１５７】

（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。

【０１５８】

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。

【０１５９】

【数３５】

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

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

【０１６０】

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。

【０１６１】

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。

【０１６２】

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。

【０１６３】

【数３６】

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

【０１６４】

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。

【０１６５】

【数３７】

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

【０１６６】

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。

【０１６７】

【数３８】

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

【０１６８】

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。

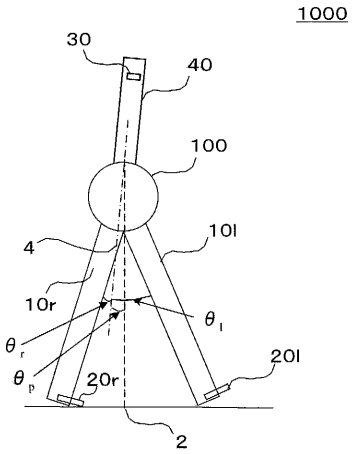
【０１６９】

【数３９】

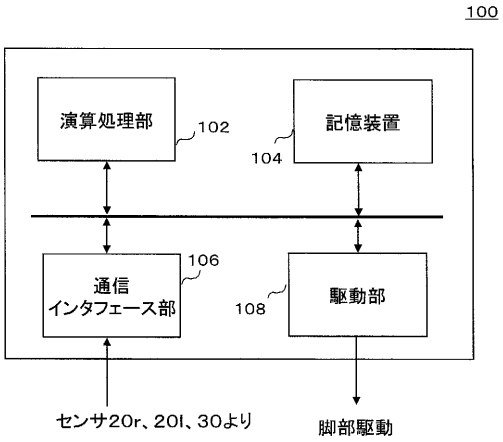
$$U_j(t) = \mu_j(t) + \sigma_j n_j(t)$$

【０１７０】
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の正規分布を表わす。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なわれる。
【０１７１】
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足歩行移動体を提供することができる。
【０１７２】
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】
【０１７３】
【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の行な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。
【符号の説明】
【０１７４】
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。

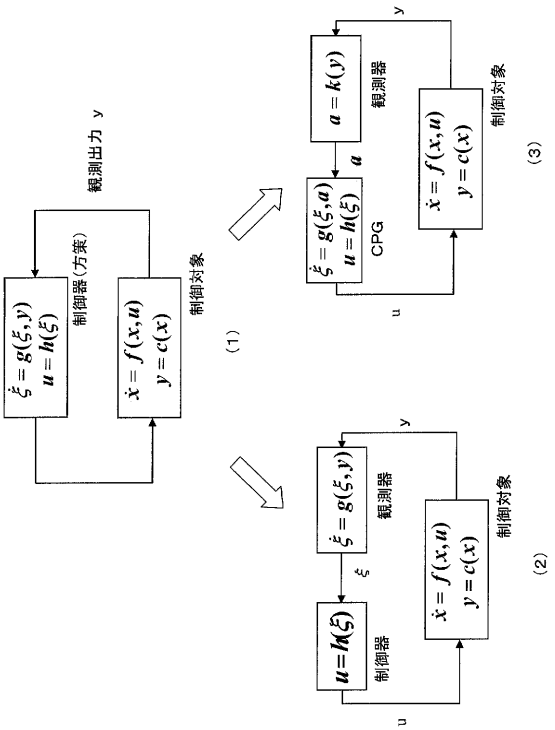
【図１】


1000
30
40
100
101
4
10r
θ l
θ r
θ p
20r
201
2

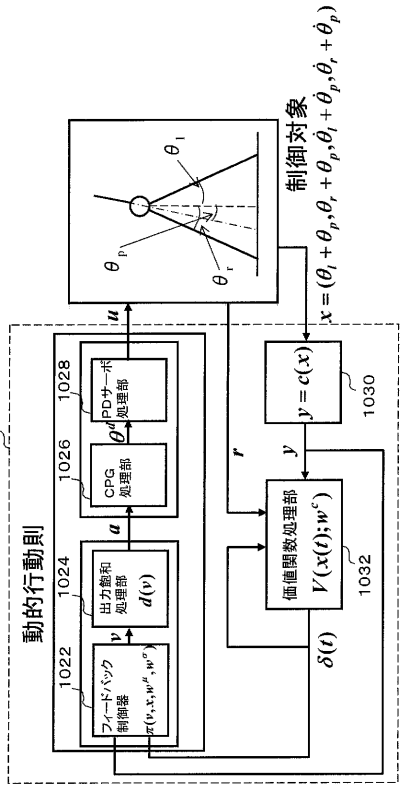
【図２】


100
演算処理部
102
記憶装置
104
通信
インタフェース部
106
駆動部
108
センサ20r、20l、30より
脚部駆動

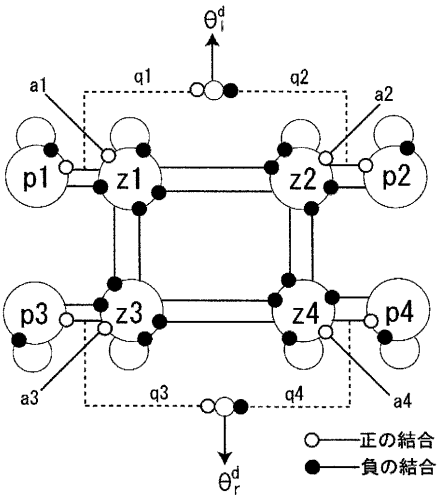
【図３】


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

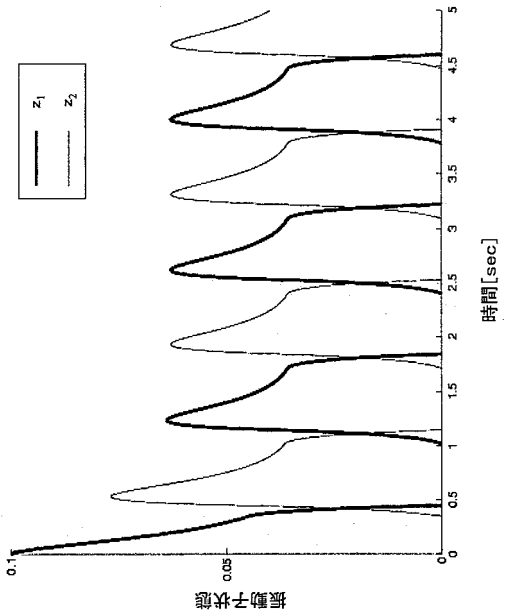
【図４】


制御対象
100
動的行動則
PDサーボ
処理部
CPG
処理部
出力飽和
処理部
フィードバック
制御器
価値関数処理部
1022
1024
1026
1028
1030
1032

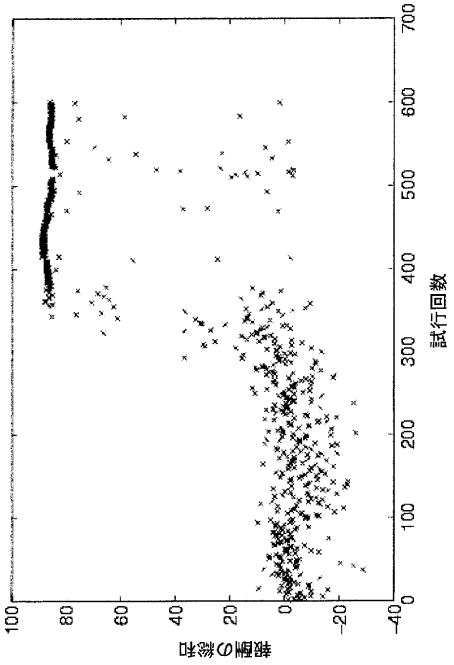
【図５】


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

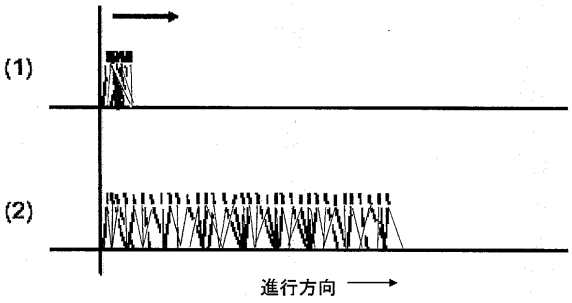
【図６】


z_1
z_2
時間[sec]
振動子状態

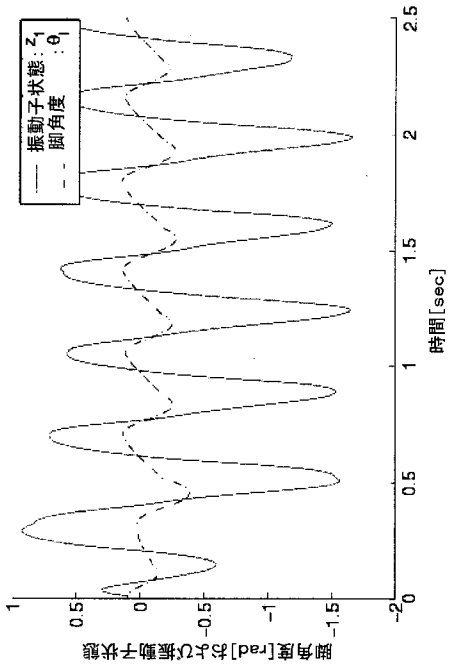
【図７】


報酬の総和
試行回数

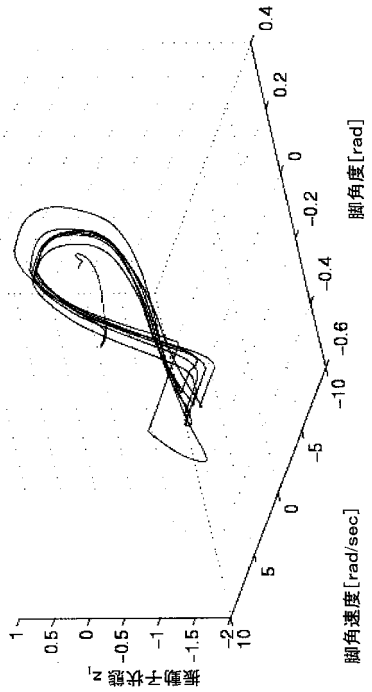
【図８】


(1)
(2)
進行方向

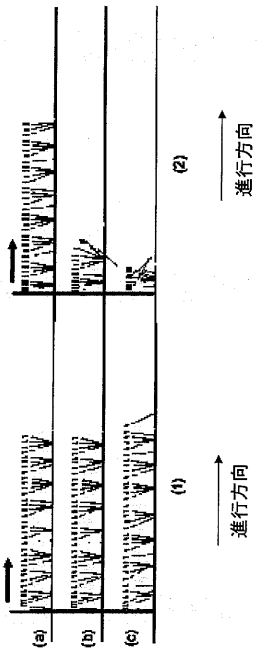
【図９】



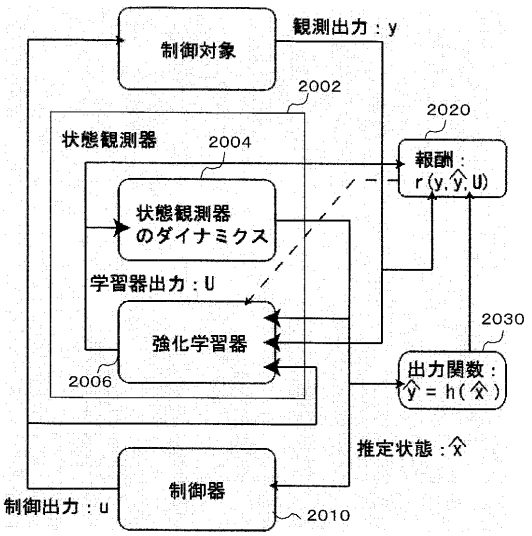
【図１０】



【図１１】



【図１２】



フロントページの続き

(74)代理人 100098316
弁理士 野田 久登
(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

審査官 植村 森平

(56)参考文献 中村泰、石井信、佐藤正樹，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，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 １／００－２１／０２
Ｇ０６Ｎ ３／００
ＣｉＮｉｉ
ＪＳＴＰｌｕｓ（ＪＤｒｅａｍＩＩ）