原著論文

身体と環境の特性を利用した状態-行動空間の抽象化 - 強化学習を用いた自律ヘビ型ロボットへの適用 -[†]

伊藤 一之^{*1}・高山 明宏^{*2}

本論文では、強化学習を用いたヘビ型ロボットの制御について考え、その際に発生する、「学習時間の長さの問題」および、「汎化能力の欠如の問題」の二つの問題を解決することを目的とする。これらの問題を解決するため、ロボットの身体および環境の性質に注目し、身体と環境との相互作用によって、状態・行動空間を実時間に抽象化する枠組みを提案する。この抽象化によって、探索空間の大きさを実学習が可能な大きさまで縮退させるとともに、学習器に汎化能力を付加する。

提案手法の有用性を確認するため、実際にロボットを製作するとともに、製作した実機を用いて学習を行った。その結果、50試行程度の少ない試行回数で学習が完了するとともに、獲得された政策は、異なる環境に対しても再学習を行うことなく適用可能であった。これにより、「学習時間の長さの問題」および、「汎化能力の欠如の問題」の二つの問題が提案手法によって解決されていることが確認された。

キーワード：強化学習、ヘビ型ロボット、状態行動空間の抽象化、汎化、次元の呪い

1. はじめに

近年、ロボットの活躍が期待される範囲は、工場などの既知環境から、災害救助活動、宇宙開発など、未知で複雑な環境へと広がってきている。そして、このような需要に対応し、複雑な未知環境において適応的に振舞うことが可能なロボットが注目され、さまざまなロボットの開発が行われている。

ヘビ型ロボットは、その有力な候補の一つである[1-4]。ヘビ型ロボットは、多数のリンクを直列に結合して構成され、小型ロボットのように狭い空間へ進入可能であるとともに、大型ロボットにも匹敵する高い走破性能を有し、小型ロボットと大型ロボットの長所を併せ持つ。

このヘビ型ロボットに、未知環境に対して自律的に適応可能な能力を付加することで、災害救助活動をはじめとするさまざまな状況で活躍するロボットの実現が可能となると考えられる。

一方、ロボットの自律的制御法として強化学習が注目されている。強化学習の枠組みでは、ロボットは、

教師信号を必要とせず、試行錯誤を通して有効な振る舞いを自律的に獲得することができる[5-12]。したがって、ヘビ型ロボットの制御に強化学習を適用することができれば、複雑な未知環境において適応的に振舞うことが可能なロボットを実現することが可能であると考えられる。

しかし、ヘビ型ロボットの制御に従来の強化学習を適用した場合には、ロボットの自由度の多さから、次元の呪いにより探索空間の大きさが膨大となり、学習が実用に耐えうる時間内に完了しないという問題が生じる。また、通常、強化学習は汎化能力が乏しく、環境が変化するたびに再学習が必要となるため、変動を伴う実環境への適用が困難であるという問題も有する。

本論文では、これらの問題の原因が身体および環境の性質を無視していることにありと考える。ロボットの身体の構成を見直すことでこれらの問題の解決を図る。そして、我々のヘビ型ロボットの制御に関する従来研究[13]を基に、環境に適した身体を設計し、この身体と環境との相互作用を用いて状態行動空間を抽象化する枠組みを提案する。実際に、提案した枠組みを基にロボットを製作し、実機を用いた学習により提案した枠組みの有用性を示す。

2. 環境とタスク

図1に、ロボットを示す。本論文では、従来研究と比較するため、ロボットは文献[13]において用いら

[†] Abstraction of state-action space by utilizing properties of the body and the environment -Application to an autonomous snake-like robot controlled by reinforcement learning-
Kazuyuki ITO and Akihiro TAKAYAMA

^{*1} 法政大学理工学部電気電子工学科
Department of Electrical and Electronics Engineering, Faculty of Science and Engineering, Hosei University

^{*2} ダイハツ工業株式会社
DAIHATSU MOTOR CO., LTD.

れたものと同様のものとする。文献[13]では、設計者が制御則を事前にロボットに埋め込んでいたが、これに対し、本論文では、強化学習により制御則を獲得することを目的とする。また、環境には、図2に示す2種類の環境、および、障害物の存在しない環境の計3種類の環境を用意し、提案する学習手法の汎化能力を検証する。

タスクの目的は、図2に示した環境において任意の場所から光源へ向かうための振る舞いを学習することである。ただし、障害物には接触してもよいこととする。障害物の間隔は最も狭い所であっても、ヘビ型ロボットの幅の1.5倍より大きいとする。また、光源は十分に高い位置にあり、障害物により光は遮られないものとする。障害物として、木の棒、自由に回転するプラスチックの筒、スポンジを巻いた木の棒の3種

類を用い、また、設置方法を固定と半固定の2種類とする。プラスチックの筒には、自由に回転する機構が付加されており、円周方向への摩擦が十分に小さい。スポンジで覆われた棒は、棒を中心にスポンジが収縮する範囲では、非線形ばね特性を持ち、その範囲を超えようとする場合には、速度拘束となる。さらに、これらのうちいくつかは、半固定状態であるため、ある程度の力が加わることで固定が外れる。この場合、障害物が外れる閾値となる力の大きさは試行ごとに異なり、事前に知ることはできない。また、ロボットは、障害物の種類や配置に関する事前情報を一切持たないこととする。

3. 従来研究と本研究の指針

本章では、学習時間の長さの問題と汎化能力の欠如の問題の二つの観点から従来研究を整理し、本論文で解決すべき問題を明らかにするとともに、本論文での解決指針を示す。

3.1 学習時間の長さの問題と汎化能力の欠如の問題

一般的に、強化学習ではロボットの自由度および環境の複雑さの増加に伴い、探索空間が指数的に増加し、学習に必要な試行回数も指数的に増加する。図1に示したロボットの場合には、関節の数は5であり、各関節の角度を10の領域に分割して状態とすれば、ロボットの関節だけでも状態数は 10^5 となる。学習にはこれに加え、環境を識別するための状態も必要となる。したがって、関節数が5程度のロボットであっても、実学習は事実上不可能であることがわかる。

さらに、最も基本的な強化学習では、政策は離散化された状態から離散化された行動への写像として保持され、学習段階においては、実際に経験した部分のみが更新されていく。したがって、未経験の状態に対応する政策は初期値のまま更新されておらず、初めて遭遇する状態が既学習の状態と非常に近い状態であったとしても、既学習の政策を活用することはできない。これは、学習効率を悪化させるだけでなく、強化学習が汎化能力を有していないことを意味する。

これらの問題に対して、状態・行動空間をさまざまな関数で近似し連続化することで汎化を行う方法が研究されてきている[10-12]。しかし、一般に、環境およびロボットが複雑になるにつれ、汎化を行うための処理も増加する。したがって、現状では、汎化を行う処理時間と強化学習を行う時間を合わせた学習時間を考えた場合には、複雑な実世界で行動する多自由度ロボットを実時間で学習させることは困難である。

したがって、学習時間の長さの問題および汎化能力

図1 snake-like robot

a) Environment for learning

b) Environment for generalization

図2 Environments

の欠如の問題の二つの問題を同時に解決するためには、環境および身体の複雑さによらず、実時間に状態・行動空間を汎化することが必要とされる。

本論文では、汎化を行うための処理をコンピュータ上で行うのではなく、環境と身体の性質を利用して実世界で行うことでこれらの問題を解決する。

3.2 未知環境における普遍性

前節で述べたように、実世界で活動する多自由度ロボットに強化学習を適用することは非常に困難である。しかし、その一方で、人間をはじめとする動物に目を向けてみると、試行錯誤を通して学習しているにもかかわらず、非常に少ない試行回数で汎化能力の高い学習を行っていることが知られている。このロボットの学習と生物の学習の本質的な違いについては未だ十分に説明されていないが、近年、さまざまな分野において実世界の性質が動物の適応的な行動を実現する上で重要な役割を果たしていることが指摘されてきている[14-22]。

例えば、生物は、種によって生息する範囲が異なり、生息可能な範囲も限られていることが知られている。生物は、環境に適した身体を持ち環境のさまざまな性質や資源を利用することで生存している。このある生物が個体群を維持していくために必要とする環境条件と資源をまとめたものは生態学的ニッチと呼ばれ[23]、文献[14]では、ロボットを設計する際のニッチの重要性が指摘されている。

この生態学的ニッチの考え方に従えば、生物にとつての未知環境とは、初めて遭遇する「ある性質をもった同種の環境」である。したがって、生物は、この環境側に存在する性質や条件を利用することで学習効率を改善している可能性がある。

しかし、従来の強化学習では、学習対象となるタスクは理想化され、何らかの確率過程としてモデル化されてきた。そして、この理想化されたモデルを対象としてアルゴリズムの設計が行われてきた。この方針のもとでは、実世界のさまざまな性質は無視されており、これらを学習に利用することは不可能であった。

本論文では、従来手法において「学習時間の長さの問題」および「汎化能力の欠如の問題」を解決することが困難である原因が、この実世界の性質を無視していることにありと考える。そして、実世界の性質として、環境と身体の特性を利用して汎化を行う枠組みを提案し、これらの問題を解決する。

なお、生物は、環境の特性を利用可能な身体を、進化の過程で獲得していると考えられるが、本論文では、進化に代り設計者が環境の特性を利用可能な身体

を設計することとする。

4. 提案手法

筆者らは、先行研究において制御系の制御量を実世界の性質を利用して抽象化する枠組みの研究を行ってきた[13]。本章では、これを発展させ、ロボットの身体の機構を環境の性質に合わせて設計することで、身体を情報処理系として利用することを可能とし、これを用いて状態・行動空間を抽象化する。

4.1 概要

図3に、提案する枠組みを示す。

提案する枠組みでは、学習器は、汎化モジュールと学習モジュールから構成されており、学習モジュールが汎化モジュールに包み込まれる形となっている。学習モジュールは通常の強化学習を行うモジュールであり、コンピュータ上にソフトウェアとして実現される。一方、汎化モジュールは、ソフトウェアではなく、身体の機構的な特性を用いて実現される。

図3の左側から右側に向かって情報の流れを説明する。まず、汎化モジュールは実世界と相互作用することにより実世界の多数の状態を抽象化し少数の汎化された状態に縮退させる。そして、これを学習モジュールに渡す。学習モジュールは、この抽象化された状態および行動を使って政策を表現し、強化学習のアルゴリズムに沿って政策を改善していく。強化学習において選択された行動は、汎化モジュールへと渡され、汎化モジュールは実世界と相互作用することでこの行動を具象化し、抽象的な行動が、多関節の運動などの複雑な動作となって実世界へ帰される。

この汎化モジュールの機能によって、学習モジュールが扱う情報は抽象化され、この抽象化された情報を用いて状態・行動空間を構成することで、探索空間の次元が低減される。これにより、学習に必要な試行回数が減少し、実学習が可能となる。

ここで重要な点は、状態および行動の抽象化・具象化はすべて汎化モジュールにより行われており、強化学習のアルゴリズムとは独立である点である。汎化モ

図3 Passive Intelligence and Active Intelligence

ジュールにおける処理は、実世界そのものを用いて並列的に行われるため、環境およびロボットの複雑さの増加により処理しなければならない情報量が指数的に増加したとしても、処理時間は増加せず、抽象化および具象化の処理は、情報量に関係なく必ず実時間に完了する[13]。これにより、環境およびロボットの複雑さに関係なく、実時間に探索空間の次元を落とすことが可能となり、従来の強化学習のアルゴリズムを用いたとしても、実学習が可能となる。

次節では、第2章で述べたタスクに対し、実際に汎化モジュールを構成し、環境と身体が状態行動空間の抽象化にどのように利用されているかを述べる。

4.2 行動の抽象化と具象化

行動を抽象化するために利用する環境の性質として、運動の法則をはじめとする、力学的性質を用いる。

図4にヘビ型ロボットのために設計した運動系および、適応的な運動の実現方法を示す。ロボットの各関節は、すべて受動関節であり、関節を扶むようにリンク間が2本のゴムでつながれている。また、リンクを囲むように2本のワイヤーが張り巡らされ、ワイヤーの長さは、最後尾に取り付けられたモータにより制御される。ロボットの下部にはクローラが取り付けられている。

図4に示すように、クローラにより推進したロボットは、障害物に接触することで反力を受け、受動関節が運動する。これに伴いワイヤーが引かれ、ワイヤーの張力により他の受動関節が運動し、障害物を回避しつつ進行方向を保とうとする振る舞いが受動的に実現される。これにより、左右のワイヤーの長さが等しい場合には、進行方向を保とうとする振る舞いが発現する(図5a)。

また、後尾のモータを回転させ、例えば、左のワイヤーを短くすることで、障害物を回避しながら左に旋回する振る舞いが発現する(図5b)。つまり、ロボットの各関節の運動は、障害物との相互作用により自動的に生成されるため、これらを学習において考慮する必要はなく、学習では、ロボットの巨視的な進行方向のみを考慮すればよい。この、ロボットの巨視的な進行方向は、ワイヤーの長さの差という一次元の値を変えることで操作することができる。ここで、この左右のワイヤーの長さの差を変化させることを行動とすれば、関節の数と同じ大きさの次元の行動空間を、一次元に抽象化することができる。また、逆の見方をすれば、このロボットの機構は、1次元の行動(進行方向)を多次元の行動(各関節の運動)へと具象化する装置であると解釈することができる。ここで、ロボットの運

図4 Passive intelligence for actuating the snake-like robot

図5 Adaptive behavior

動は、関節の数や、接触している障害物の数に関係なく、必ず実時間に行われるため、行動の具象化も必ず実時間に行われる。

以上により、上述の環境および身体の利用可能な環境であれば、コンピュータによる処理を一切必要とすることなく、行動の抽象化および具象化が可能となる。

4.3 状態の抽象化

本項では、状態をロボットの進行方向を表す一次元の状態とロボットから見た光源の方向を表す一次元の状態の計二次元の状態へと抽象化する。

最初に、進行方向を表す状態について述べる。通常、ロボットの進行方向は各リンクの状態および障害物との接触状態によって決まるが、前項で述べたように、ロボットの機構を適切に設定することで、図4のワイヤーの長さの差という一次元の値に抽象化することが可能となる。したがって、図4の機構により、状態も行動と同様に、「現在のワイヤーの長さの差」を進行方向を表す状態として用いることができる。

次に、ロボットから見た光源の方向を表す状態について述べる。抽象化には、CdS素子の指向性という実世界の性質を用いる。図6に方向を取得するための仕

図6 Passive intelligence using CdS cells

組みを示す。

CdS素子として浜松ホトニクス製の可視光導電素子(図6a))を用いる。この素子は、光量に応じて抵抗値が変化する素子であり、光が強いほど低い抵抗値を示す。CdSは、4つを一組として図6b)のように半円弧状に配置され、さらに、各リンクに一組ずつ、図6c)のように配置される。

式(1)により各素子が受光した光の強さの重心に相当する値を算出し、これをロボットから見た光源の方向とする。ここで、 N はリンク数を表す。式(1)の領域は $0 < x < 3$ であり、ロボットが光源の方向へ直進しているとき1.5を返す。式(1)の計算は、アナログ電子回路により並列的に処理することとする。

以上により、行動一次元、状態二次元の合計三次元まで、状態・行動空間を低減することが可能となる。

$$x = \frac{\sum_{i=1}^N (1/R_{i2}) + 2 \sum_{i=1}^N (1/R_{i3}) + 3 \sum_{i=1}^N (1/R_{i4})}{\sum_{j=1}^4 \sum_{i=1}^N (1/R_{ij})} \quad (1)$$

4.4 学習モジュール

本論文で提案する手法は、必ずしも学習方法を限定するものではない。ここでは、一つの実例として、Q学習[24]を用いた例を示す。Q学習は、状態・行動を離散化して表現する代表的な手法であり、Q学習のアルゴリズム自体は汎化能力を持たず、また、学習速度を改善するような仕組みを持たない。したがって、Q学習を用いることで、汎化能力および学習時間の短縮が提案手法によるものであることを容易に確かめることができる。以下に、本論文でのQ学習の設定を述べる。

更新式には、式(2)で示される典型的なQ学習の更新式を用いる。

$$Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha \{ r(s, a) + \gamma \max_{a'} Q(s', a') \} \quad (2)$$

ここで、 s は状態、 a は行動、 s' は s において a を実行することによって遷移した状態、 r は報酬、 α は学習率、 γ は割引率を表す。

次に、状態および行動の構成について述べる。状態は、式(1)の計算を行う回路の出力、および、ワイヤーの長さを制御するプーリーの角度(ワイヤーの長さの差に相当)とし、それぞれ、表1および、表2で示すように5つの領域に分割する。したがって、離散化された状態の個数は25個である。

行動は、ワイヤーの長さを制御するプーリーの回転角度とし、行動1：-25度回転、行動2：現在の角度を保持、行動3：+25度回転の3種類とする。ただし、プーリーの角度の最大値は左右にそれぞれ50度であり、これを超える方向へ回転を行う行動が選択された場合には、回転は行わないものとする。

また、クローラを制御するための行動は設けず、クローラのモータには、常に一定の電圧を印加し、ロボットを前進させる方向へ回転させるものとする。

以上の設定により、状態・行動空間は、三次元空間で合計75領域となる。これは、ロボットのリンク数が6、さらに環境にさまざまな障害物が存在することを考慮すると非常に少ない数であると言え、このことから、抽象化により状態・行動空間が縮退されていることが確かめられる。

報酬は、光源の方向が3の状態にある時(光源に向かっていているとき)、正の報酬100を与える。また、光源の方向が測定範囲外に遷移したとき(光源を見失ったとき)は、負の報酬-100を与え、その試行は打ち切り、再び初期状態に戻して次試行へと移る。ロボットの初期位置は、図7に直線で示した23種類のなかから無作為に選択するものとする。

表1 State of the light direction

state	direction
1	$0 \leq x < 1.35$
2	$1.35 \leq x < 1.47$
3	$1.47 \leq x < 1.53$
4	$1.53 \leq x < 1.65$
5	$1.65 \leq x < 3$

表2 State of the body

state	1	2	3	4	5
angle[deg]	-50	-25	0	25	50

図7 Initial position

5. 実験

実機を用いて学習を行い、抽象化された状態・行動空間を使用して学習が可能であることを示す。また、得られた政策を未知の環境に対して適用することで、提案手法の汎化能力の有無を検証する。

5.1 実機を用いた学習

障害物の存在しない環境と障害物の存在する環境(図2a))の二つの環境でそれぞれ実機を用いて学習を行い、それぞれの環境において、光源へ向かう振る舞いが獲得されることを確認する。

学習率を0.2、割引率を0.5として学習を行う。

行動の選択およびQ値の更新は2.5秒間隔で行い、この間に、プーリーは選択された行動にしたがって、目標の角度まで回転する。その後プーリーの角度は、次の行動が選択されるまで保持される。

行動の選択には、 ϵ -greedy法を用い、10%の確率で無作為に、90%の確率でQ値が最大となる行動を選ぶ。1試行を40秒間とし、100試行で一つの学習を終了する。これを、障害物の存在しない環境と、障害物の存在する環境(図2a))とでそれぞれ10回行い、計20の政策を獲得する。

図8に、10試行ごとの報酬獲得率(10試行中、少なくとも一回は正の報酬が与えられた試行の割合)を示す。各値は、10回の学習における平均値である。

これらの結果から、障害物の有無にかかわらず、どちらの場合にも、50試行程度の学習で、学習が収束し、光源へ向かって移動する振る舞いが獲得されることが分かる。また、この高い報酬獲得率を持った政策が獲得されたという結果から、抽象化された状

図8 Reward acquisition ratio

態・行動空間は、大きさは非常に小さいものの、このタスクを行う上では十分なものであり、抽象化が成功していることが分かる。さらに、50試行という少ない試行回数で学習が可能であるという結果は、「学習時間の長さの問題」が解決されていることを示している。

なお、報酬獲得率が1に収束していないのは、行動の選択に ϵ -greedy法を用いているためである。また、障害物の存在する環境における報酬獲得率が障害物の存在しない環境における報酬獲得率に比べて若干低くなっている原因は、障害物の存在によって、状態遷移が確率的に行われるためであると考えられる。障害物の存在しない環境においては、ロボットを右へ旋回させる行動をとった場合、確率1でロボットは右へ旋回するが、障害物が存在する場合には、障害物との接触により2.5sという限られた時間内には、右へ旋回しない場合もあり、確率は1とはならない。

5.2 汎化能力の検証

汎化能力を検証するため、以下の4種類の実験を行う。

実験1：障害物の存在しない環境における学習で獲得された政策を、障害物の存在する環境(図2a))へ適用

実験2：障害物の存在する環境(図2a))で獲得された政策を、障害物の存在しない環境へ適用

実験3：障害物の存在する環境(図2a))で獲得された政策を、学習が行われた環境とは異なる環境(図2b))へ適用

実験4：障害物の存在する環境(図2a))で獲得された政策を、学習が行われた環境と同一の環境(図2a))へ適用

なお、行動の選択には、 ϵ -greedy法は用いず、最大のQ値をとる行動を選択するものとする。

表3に各30回の実験におけるタスクの成功率(光源

を中心とする半径20cmの円内に到達した場合を成功とした)を、図9に、実現された振る舞いの一例として、実験1の結果を示す。

実験の結果、4つのすべての実験において、図9に示すような、光源へ向かって推進する振る舞いが観測され、成功率も非常に高いものであった。

また、表3から、学習を行った環境と同一の環境へ適用した場合(実験4)の成功率と、学習を行った環境とは異なる環境へ適用した場合(実験1, 2, 3)の成功率が、ほぼ同じであることが分かる。これらの結果は、獲得された政策が、学習環境とは異なる未知環境に対して、一切の追加学習を必要とせず適用可能であることを示しており、提案手法により「汎化能力の欠如の問題」が解決されていることが分かる。ここで、今回獲得された政策は、従来のQ学習と同様に、離散

化された状態と行動からなるQ値を用いて表現されているものであり、このQ値自体が汎化能力を持つわけではないという点が重要である。汎化能力は、提案手法によって状態と行動を抽象化することによって実現されている。

なお、表3に示したように、障害物の存在する環境において、光源へ到達できない振る舞いが数回観測されたが、これは、光源付近で障害物と接触したことによる。障害物との接触により光源とは逆方向へ大きく旋回してしまい、この後、光源方向へ旋回しなおしたものの間に合わず、光源へ到達することが不可能であった。しかし、この場合でも、光源へ向かおうとする振る舞いは観測されており、学習により獲得された政策自体は環境によらず正常に機能していると考えられる。

6. 結論

本論文では、強化学習による実ヘビ型ロボットの制御を目的とし、その際発生する問題として、「学習時間の長さの問題」および「汎化能力の欠如の問題」の二つの問題について考えた。そして、環境と身体の特性に注目し、これらを同時に解決するための枠組みを提案すると共に、実際に状態・行動空間を抽象化可能な身体の機構を設計した。提案手法の有用性を確認するため、実ロボットを用いて学習を行い、実学習が可能な程度まで状態行動空間の次元が低減されていること、および、獲得された政策が環境の変動に対して高い頑健性を持つことを実験により示した。

本論文の重要な主張は、従来の問題がアルゴリズムの改良によって解決されるのではなく、実世界のさまざまな性質を用いることによって解決されるという点である。本論文の結果は、ヘビ型ロボットという一具体例に留まっており、提案する枠組みの一般化には至っていないが、本論文で取り上げた問題は、実ロボットに強化学習を適用する上での非常に大きな問題であり、これらを解決する一つの指針を示したことは、今後、未知環境で働くロボットの制御系を構築する上で非常に有用であると考えられる。

今後は、環境を3次元化し、より現実の不正地へ近づけるとともに、光源の代わりに人間の呼吸を目標とするなど、タスクを災害現場における人命探査などに応用可能な形へ変更してゆく予定である。

参考文献

- [1] Masayuki Arai, Toshio Takayama, Shigeo Hirose, Development of Souryu-III: Connected Crawler Vehicle for Inspection inside Narrow and Winding

表3 Success rate

Experiment	Success rate
1	26 / 30
2	30 / 30
3	24 / 30
4	26 / 30

図9 Behavior

- Spaces, Proc. of Int. Conf. Intelligent Robots and Systems, pp.52-57, 2004
- [2] K. L. Paap, T. Christaller, and F. Kirchner, A robot snake to inspect broken buildings, in Proc. of Int. Conf. Intelligent Robots and Systems, pp.2079-2082, 2000.
- [3] 大須賀公一, ヘビ型ロボットによる狭空間探索, 日本ロボット学会誌, Vol.22, No.5, pp.554-557, 2004.
- [4] K. Ito and R. Murai, Snake-like robot for rescue operation, -Proposal of simple adaptive mechanism designed for ease of use-, Advanced Robotics, Vol.22, No.6-7, pp.771-785, 2008.
- [5] R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction, MIT Press, 1998.
- [6] 谷口忠大, 榎本哲夫, 報酬設計を通じた社会的相互作用による行為概念群の構築: シェマ理論に基づいた累積的強化学習, 知能と情報, Vol.18, No.4, pp.629-640, 2006.
- [7] 山村忠義, 馬野元秀, 瀬田和久, 段階的な視覚をもつエージェントにおける強化学習について, 知能と情報, Vol.18, No.4, pp.561-570, 2006.
- [8] 木村元, 荒牧岳志, 小林重信: “重み付けされた複数の正規分布を用いた政策表現最適行動変化に追従できる実時間強化学習と環状ロボットへの適用”, 人工知能学会誌, Vol.18, No.6, pp.316-324, 2003.
- [9] 内部英治, 銅谷賢治: “複数報酬のもとでの階層強化学習”, 日本ロボット学会誌, Vol.22, No.1, pp.120-129, 2004.
- [10] 近藤敏之, 伊藤宏司: “進化的 recruitment 戦略を用いた強化学習による自律移動ロボットの制御器設計”, 計測自動制御学会論文集, Vol.39, No.9, pp.857-864, 2003.
- [11] 堀内匡, 藤野昭典, 片井修, 榎本哲夫: “連続値出力を扱うファジィ内挿型 Q-Learning の提案”, 計測自動制御学会論文集, Vol.35, No.2, pp.271-279, 1999.
- [12] 石井信, 佐藤雅昭: “オンライン EM アルゴリズムを用いた強化学習法”, 電子情報通信学会技術研究報告, NC, Vol.98, No.577, pp.41-48, 1999.
- [13] 伊藤一之, 福森嘉孝: “知覚量に基づく制御系設計-ヘビ型ロボットの方向の知覚量を用いたフィードバック制御-”, 計測自動制御学会論文集, Vol.42, No.4, pp.436-445, 2006.
- [14] R. Pfeifer, C. Scheier 著, 石黒章夫, 小林宏, 細田耕監訳: 知の創成, 共立出版, 2001.
- [15] 松野文俊: “生物に学ぶ機械と制御-ダイナミクスベーストインテリジェンス-”, 計測と制御, Vol.42, No.9, pp.699-703, 2003.
- [16] 佐々木正人: “アフォーダンス入門-若きロボット研究者との Q & A-”, 日本ロボット学会誌, Vol.24, No.7, pp.791-796, 2006.
- [17] R. Pfeifer, F. Iida and G. Gomez: “Designing Intelligent Robots - On the Implication of Embodiment-”, 日本ロボット学会誌, Vol.24, No.7, pp.791-796, 2006.
- [18] 有本卓: “巧みさの演出: ダイナミクスベースト制御”, 日本ロボット学会誌, Vol.24, No.7, pp.791-796, 2006.
- [19] 大須賀公一: “ダイナミクスベースト制御の「こころ」”, 日本ロボット学会誌, Vol.24, No.7, pp.791-796, 2006.
- [20] 石黒章夫: “知の基盤としてのしぶとさの創成”, 日本ロボット学会誌, Vol.24, No.7, pp.791-796, 2006.
- [21] 古山宣洋: “運動を導く知覚システム”, 日本ロボット学会誌, Vol.24, No.7, pp.791-796, 2006.
- [22] 伊藤一之, 大須賀公一, 石黒章夫, 古山宣洋: “実世界の性質を利用した知覚と制御”, 日本ロボット学会誌, Vol.24, No.7, pp.791-796, 2006.
- [23] M. Begon, J. L. Harper, C. R. Townsend 著, 堀道雄監訳: 生態学-個体・個体群・群集の科学, 京都大学学術出版会, 2003.
- [24] C. J. C. H. Watkins, P. Dayan, Technical note Q-learning, Machine Learning, 8: 279-292, 1992.

(2008年8月17日 受付)

(2009年1月21日 採録)

[問い合わせ先]

〒184-8584 東京都小金井市梶野町3-7-2

法政大学 理工学部 電気電子工学科

伊藤 一之

TEL: 042-387-6093

FAX: 042-387-6381

E-mail: ito@hosei.ac.jp

著者紹介

いとう かずゆき
伊藤 一之 [正会員]

2002年3月 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博士後期課程修了, 同年4月岡山大学工学部システム工学科助手, 2005年4月法政大学工学部システム制御工学科専任講師, 2009年4月同大学理工学部電気電子工学科准教授となり現在に至る, 2003年日本ロボット学会研究奨励賞, 2004年船井情報科学奨励賞などを受賞, 主に知能ロボットに関する研究に従事, 計測自動制御学会, 日本ロボット学会などの会員, 博士(工学).

たかやま あきひろ
高山 明宏 [非会員]

2008年3月 法政大学大学院工学研究科システム工学専攻修士課程修了, 同年4月 ダイハツ工業株式会社入社現在に至る.

**Abstraction of state - action space by utilizing properties of the body and the environment
- Application to an autonomous snake - like robot controlled by reinforcement learning -
by**

Kazuyuki ITO and Akihiro TAKAYAMA

Abstract :

In this paper we consider autonomous control of a real snake-like robot using reinforcement learning. We focus on curse of dimensionality and lack of generality, and point out that the causes of these problems are not in learning algorithm but in neglect of properties of the real world. To solve these problems we propose new framework in which the body of the robot abstracts general meanings by interacting with environment. We apply the proposed framework to control of a real snake-like robot and confirm that the two problems are solved.

Keywords : reinforcement learning, snake-like robot, abstraction of state-action space, general policy, curse of dimensionality

Contact Address : Kazuyuki ITO

Department of Electrical and Electronics Engineering, Faculty of Science and Engineering, Hosei University

3-7-2, Kajino-cho, Koganei, Tokyo 184-8584, JAPAN

TEL : 042-387-6093

FAX : 042-387-6381

E-mail : ito@hosei.ac.jp