特　集

強化学習の実ロボットへの応用とその課題

Research Issues on Real Robot Reinforcement Learning

浅田　　稔*
Minoru Asada

* 大阪大学大学院工学研究科
Graduate School of Engineering, Osaka University

19YY年MM月DD日　受理

Keywords: Reinforcement Learning, Real Robots, State-Action Space, Sensor Space, Motor Space

1. はじめに

動的に変化し続ける実環境において真に知的に振る舞うシステムは，環境との相互作用を通して，何らかの形でシステム内部に目的行動を生成するための環境の記述を構築する必要があり，学習過程が欠かせない．このようなシステムの代表は実時間での意思決定問題を扱わなければならないロボットであろう．ロボットの学習法としては，種々のアプローチが提案されているが，最近，反射的かつ適応的な行動を獲得できる手法として，強化学習が注目を浴びている [CM93b](強化学習をはじめとするロボット学習の概要に関しては、他の文献，例えば、[Mah96] などを参照されたい．)．

本稿では，実際のロボットへ適用する際の問題点として,「状態・行動空間の構成」,「複雑なタスクへの対応」の二つの問題を指摘する．前者は、ロボットの物理的なセンサやアクチュエータの空間が、状態や行動と必ずしも一致しないために生じる様々な問題を指す．後者は、学習の加速化及び前者の問題を含めたスケールアップ問題を指す．以下では、まず強化学習の基本的な枠組を述べた後，状態・行動空間構成問題を中心に，これら二つの問題の関係及び課題を実例を通して述べ，最後にまとめる．

2. 強化学習の枠組

強化という言葉は元々行動心理学の用語として用いられていた．アメリカの行動心理学者の代表であるスキナーは，人間をはじめとする動物の行動を説明するための基本原理として「強化」による行動原理を唱え，種々の動物行動実験を行った．代表的な実験として，スキナーボックスによる鼠の行動実験を例に強化学習の基本的枠組みと，その問題点を簡単に説明しよう．

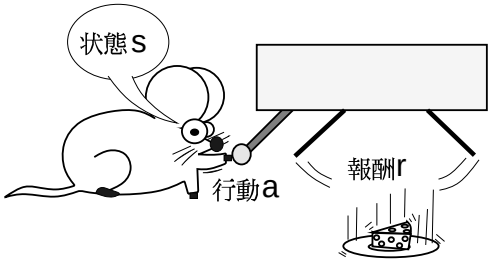


図1 スキナーの鼠箱

スキナーボックスの実験では，鼠を箱の中に入れ，その中にあるレバーを鼠がたまたま押すと，餌がもらえる実験で，一旦レバー押しを憶えると何回もレバーを押し続ける行動をとる (図1参照)．このときレバーを押す行為に正の強化 (餌，報酬，価値などなど) が与えられる．強化学習は，これを確率的動的計画法の枠組で定式化したものである．

鼠は箱の中で，どこにいたり，レバーがどのように見えるかなどの状態 ($s \in \boldsymbol{S}$:状態集合) が分かり，前に進んだり，レバーを押すなどの行動 ($a \in \boldsymbol{A}$:行動集合) をとることができる．このとき，環境は厳密にはマルコフ過程としてモデル化され，現在の状態と鼠がとった行動により確率的に (うまく見えなかったり，脚を滑べらしたりするかもしれないので) 次の状態 ($s' \in \boldsymbol{S}$) 遷移する．その結果報酬 (r：例えばチーズ) が与えられる．状態遷移が既知であれば通常の動的計画法 (以下,

DP と略記) の枠組で最適行動が得られるが，未知のとき環境内で試行錯誤しながら，状態遷移と最適行動を推定しなければならない．これが確率的DP とか逐次的DP などと呼ばれる結縁である．最も良く利用される強化学習法として**Q 学習** [WD92] が有名で，状態 s で行動 a をとる行動価値関数 $Q(s,a)$ は，試行錯誤により，次式で更新される．

$Q(s,a) \Leftarrow$

$$(1-\alpha)Q(s,a)+\alpha(r(s,a)+\gamma \max_{a' \in \boldsymbol{A}} Q(s',a')) \quad (1)$$

ここで，s' は，次状態，α は学習率で 0 と 1 の間の値をとる．γ は，減衰率で，現在の行動が将来に渡ってどれくらい影響を及ぼすかを決めるパラメータで，0 と 1 の間の値をとり，小さい程影響が少ない．行動選択は，学習の収束時間を決める要因の一つで，一旦憶えた成功例を何回も繰り返して上達させるか，別のアプローチを未経験のところから探すかのトレードオフがある．無限の時間を費やして探索することが困難な実ロボットの観点からは前者が有利であるが，準最適解しか発見できない可能性が高くなる．これらの振る舞いの理論的な解析は，本特集号の別の解説で説明がなされているので，以下では，実ロボットへの適用の観点から二つの問題を取り上げ，それぞれについて説明する．

3. 状態・行動空間の構成問題

スキナーボックスでは，鼠は，箱の中のどこにいたり，レバーの見え方などの状態の定義を鼠自身が事前にもっていることを仮定していた．実際の環境では，これらの状態そのものをどのように定義するかが大きな問題である．すなわち，

(1) どのような情報を使えば，タスク遂行に必要かつ十分な状態空間が構成可能であるか，

(2) マルコフ性を満足する状態・行動空間をどのように構成するか．

従来の強化学習の研究では，多くがコンピュータシミュレーションによるもので，実ロボットへの適用可能性を論議しているものは少なく，ロボットの行動により状態が次状態に遷移する理想的な行動及び状態空間を構成している．しかしながら実環境で作動するセンサやアクチュエータの出力が直接，状態や行動に 1:1 に対応するとは限らない．むしろ目的に応じて，センサ空間やモータ空間を抽象化し，状態・行動空間を構成することが望まれる．このとき、センサ情報から状態へ，またモータ出力から行動への抽象化過程は，相互に依存し，鶏と卵問題に類似している (図先に行動空間をプログラマが設計し，それに基づいて状態空間を構成するものがほとんどであった．しかし，相互依存性を考慮すると，必ずしもそのような設計がうまくとは限らない．以下では，センサ出力，モータ出力と状態・行動を厳密に区別しながら，1) モータ空間を抽象化し，行動空間を規定する手法，2) 状態空間を先に設計し，それに基づき行動空間を定義する手法，3) 何らかの拘束条件の基に，状態・行動空間を構成する手法を実例を交えて説明する．

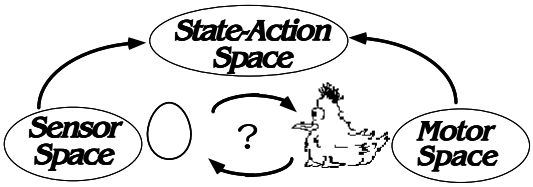


図 2 「鶏と卵」問題

3・1 行動空間の設計

先にも述べたように，物理的なセンサやアクチュエータが状態や行動に 1:1 に対応するとは限らないので，センサ情報の抽象化だけでなく，モータ出力の抽象化も考慮されなければならない．アクチュエータの物理的自由度が少なくても，モータコマンドが連続空間で与えられているとき，何らかの抽象化が必要となる．中村，浅田ら [中村 96] は，2 自由度の移動ロボットの様々なモータコマンドの実行により，ロボットに搭載されたTV カメラから観測されるフローパターンの主成分解析から，移動ロボットの物理的自由度 (直進と回転) を算出し，モータコマンドを抽象化した．これにより，以降の学習の高速化を実現した．彼らの手法は，高々2 自由度の移動ロボットを対象にしており，より多自由度の場合のセンサ情報との非線型なマッピングに対しては，今後の課題とされている．

3・2 状態空間を規定し行動空間を構成する手法

ソナーやバンパーなどの近接・局所センサでは，センサ情報の変化は，モータコマンドによる環境への働きかけと対応しており，各モータ出力が状態遷移に対応可能である．しかしながら，視覚情報のように射影を通して得られるセンサ空間では，同じ物理的動作が，画像上で異なる変化を引き起こし，正しく学習可能な状態と行動を定義することは難しい．例えば，ボールをゴールにシュートするサッカーロボットの例 [ANTH96] では，状態数を低減するためにボールやゴールの位置，大きさ，向きを粗く (左中右や大中小など) サンプルした

状態空間を用意したが，同じ状態で，同じモータコマンドを実行しても，ほとんど元の状態に戻り，学習が進まない(図3参照)．この問題は**「状態と行動のずれ問題」**と呼ばれている．これに対し，彼らは，状態が変わるまで，同じモータコマンドを出力し続け，その一連の出力を行動とすることで,「ずれ」が生じないような行動空間を再構築した．このことは，物理的に絶対的な長さが規定された行動があるのではなく，現在の状態にしたがって，長さの変化する行動が生成されることを意味し，時間の相対性をあらわしていると考えられる．しかしながら，最初に設計者が与えた状態空間がロボットにとって最適である保証はない．

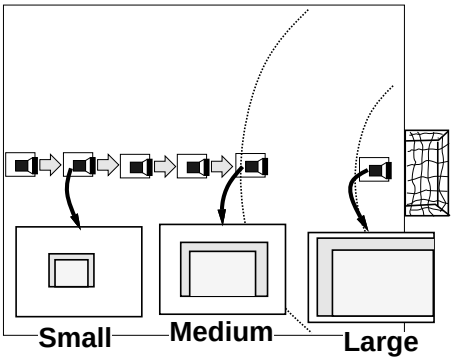


図3 「状態と行動のずれ」問題

3・3 状態と行動を同時に構成する手法

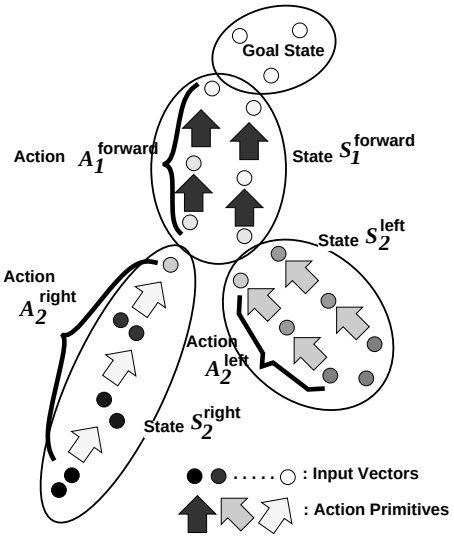


図4 状態と行動の相互規定

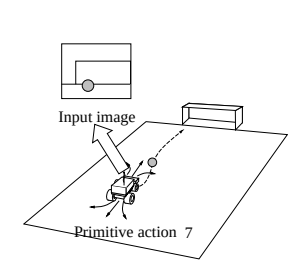
Asada et al.[ANH96] は，先の手法 [ANTH96] を発展させて，状態空間と行動空間を同時に構成する手法を示した．単位時間当たりに実行されるモータコマンドを行動要素，その結果生じる環境の変化を感知するセンサ情報を入力ベクトルとして，同一行動要素の系列でゴール状態 (もしくは既に獲得されている状態) に到達できる入力ベクトルの集合を「状態」，そのときの行動要素の系列を「行動」として定義し，実ロボットがボールをゴールにシュートするタスクに適用した．図4にその基本的な考えを示す．小円が入力ベクトルを示し，濃淡が，知覚の違いを示す．太い矢印が行動要素 (モータコマンドの種類) を示す．異なる入力ベクトルが知覚の違いに関わらず，タスク (ゴールへの状態遷移) に応じて，同じ状態にクラスタリングされている様子がわかる．

Ishiguro et el.[ISI96] は，全方位に移動可能な3自由度移動ロボットの時間的に連続する2枚の全方位画像から，ロボットの航行のための状態空間を構成した．但し，探索空間が膨大なので，教示データを基に，特徴ベクトルを直交化し，木構造の状態空間を構成した．結果得られる特徴ベクトルは，ロボットが移動中に視覚情報のどこに注視すべきかの情報も示している．

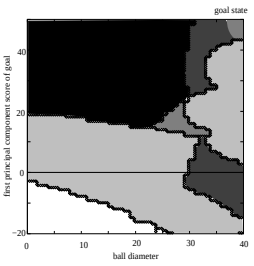
これら二つの例は、いずれもオフラインの学習である．前者では、ゴール状態からバックトラックすることで，後者では，人がゴール行動の一例を示すことで探索空間を低減し，状態空間を構成するプロセスそのものが学習過程に対応する．タスク遂行に必要な情報の取捨選択は，前者では凸包，後者では特徴空間の直交化により実現されており，センサ情報が入組んだ非線型な特徴空間の分離は困難である．また，後者では行動空間は，モータ空間に対応しており，抽象化は行われていない．また，双方とも，オフライン処理のため，分割するのみである．

オンラインの手法として，実ロボットに適用された例ではないが，ゴール状態へ導く初期コントローラ (必ずしも正しく導く必要はない) を想定し，ゴール状態とそれ以外からなる初期の状態空間を再帰的に分割していく PARTI-GAME アルゴリズ [MA95] が知られている．ゴール状態へ到達できないとき，適当に状態空間を分割し，サブゴールを生成していく．分割法に関する指針が明示的に与えておらず，ゴール状態が厳密に規定されているときは，徒に分割を繰り返す惧れがある．

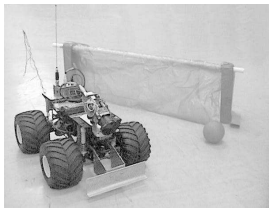
実ロボットに適用されたオンライン状態・行動空間構成法として，Takahashi et al.[TAH96] は，関数近似を分割指針とし，状態の分割だけでなく融合過程を導入することにより，無駄な再分割を防ぐ手法を提案



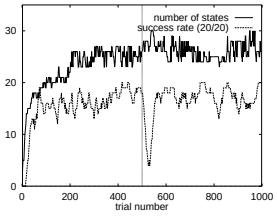
(a) タスク



(b) 状態空間



(c) 実ロボット



(d) 成功率と状態数

図5 タスクと実験結果

している．具体的なタスクとして，視覚移動ロボットがボールをゴールにシュートするタスクを考えた．(図5(a)参照)．画像から得られるのはボールの位置(x座標)と大きさ，ゴールの位置，大きさ，向きの5次元パラメータであり，行動は二つの独立した左右輪への回転指令である．学習の基本的な考え方は，以下である．

(1) 5次元の知覚空間は最初2状態(目標状態とそれ以外)からなる．

(2) 行動に関する状態変化を関数近似し，近似による状態変化予測が異なる場合か，ゴール到達に失敗した場合のみ，状態を分割または融合し，あらたな関数近似領域を推定する．これにより無駄な探索を軽減できる．

(3) 新たに分割された状態の行動価値のみを初期化し，通常の強化学習を適用する(状態数が少ないので学習時間が短い)．

(4) 行動選択にランダムネスを付加し，環境の変化に対応する．

図5(b,d)に実験結果を示す．図(d)の実線と破線はそれぞれ状態数と過去20試行の成功回数を示している．450回目にボールの大きさを2倍に変更した直後は成功率が下がったが，直ちに持ち直している様子が分かる．図(b)は分割された5次元の状態空間を2次元(ボールとゴールの大きさ)に射影したもので，右上がゴール状態を示している．入り組んだ状態が獲得された様子が分かる．状態変化が生じるまで同じモータコマンドを発生させ，その系列を行動と定義することで，行動の時間的分割を行っている．因みに1時間半ほどの実ロボット(図(c)参照)の学習時間で目的の行動が達成できた．

4. 複雑なタスクへの対応問題

学習の更新式(1)では，状態遷移毎に即座の報酬が与えることができるが，いい加減な報酬関数では，行動価値関数が多くの極値をもち，学習が収束しない．逆に，収束が保証される詳細な報酬関数が既知である事は，制御則が既知であり，学習を要しない．そこで，単純なタスクでは，最終ゴール状態のみに報酬を与え，それら以外は，ゼロ報酬とする場合が多い．スキナーボックスの例では，鼠がたまたまレバーを押すことで，餌がもらえ，レバー押し行動が強化されるが，このことは，餌(強化信号)が得られることの偶然性を期待しており，これが，遅延報酬による学習時間の増大を招いている．

この問題に対し，事前にサブタスクに分解する手法などが提案されているが，タスクがより複雑になると，互いに干渉しない独立なサブタスクへの分割が困難となる．スキナーボックスでは，箱の中にレバー以外のもの，特に餌をとる他の鼠がいる場合，その行動により自身の行動と無関係に状態が遷移する可能性がある．更に，他の鼠との協調や競合などの行動をどのように学習するかなど，前節の課題を含めて問題が複雑化し，これに対応しなければならない．

4・1 サブタスクへの分割による対応

強化学習の場合，収束するまでの学習時間は，状態空間のサイズの指数オーダーとなる[Whi91]ので，複雑なタスクに対し，直接強化学習などを利用する事は，ほとんど不可能に近い．一つのアプローチは，サブタスクに分解し，個々のサブタスクを個別に学習し，それらを統合することである．この時の問題は，

課題1サブタスク間で干渉が生じる場合，どのように対応するか，

課題2サブタスクの切り替えをどのように設計するか，

課題3さらに学習をどのように加速するか，

である．

Mahadevan and Connel[CM93a]は，ロボットの箱押し作業で，実環境での学習に多大な時間を要するので，作業を事前に「箱の発見」,「箱押し」,「スタック状

態からの回避」の3つに前もって分割し，それぞれに強化学習を適用して，学習時間の短縮化を図った．個々のサブタスクの状態空間が独立で干渉せず，時系列的に実行可能であるので，課題1，2の問題は生じない．また分割による高速化のみで，課題3には対応していない

Uchibe et al.[UAH96]は，ゴールキーパーを相手にボールをシュートするタスクで，先に獲得したシュート行動[ANTH96]と，別に獲得した回避行動を統合することを考えたが，ボールがゴールキーパーに隠されるなどの状態空間が干渉し，隠れ状態が発生する，二つの行動を切替える条件は，状況に依存し容易に決定できない，膨大な学習時間を要するなど上記の課題が全て含まれていた．そこで，彼らは，最尤推定を用いて，隠れ状態を推定・識別した(課題1)．また個別に獲得された行動価値関数の単純和を初期値として，隠れ状態近傍を集中的に探索することで，自動的に二つの行動の切替え条件を学習させると同時に学習を加速させた(課題2,3)．それでもなお，多大な学習時間を要するので，個別に学習された行動価値観数に基づいて，全空間を探索するのではなく，個々の行動が支配的である空間での再学習を行わないことにより，学習時間を約1/3に短縮した．

4・2 より複雑なタスクへの対応

これまでのタスクでは，一部の隠れ状態を除いて，センサ空間の次元が状態空間の次元と等しい場合を扱ってきた．即ち，ほとんどが静止環境を想定していた．しかし，環境が動的に変化する場合，現在のセンサ情報だけから適切な行動を決定することが困難となる．これまでこの種の問題は，強化学習の分野では部分観測マルコフ問題として定式化されてきた[LCK95]が，他のロボットを含むマルチエーエジェント環境では，問題はより深刻となる．即ち，自身の行動と直接関係なくセンサ情報が変化するので，通常のセンサ情報に直接基づいた状態空間では強化学習を実現できない．

これに対し，実ロボットへの適用はなされていないが，木構造を用いて時間的な履歴を状態として記述する手法が提案されている[McC95]．この場合，すでにシンボル化された記述がノードに対応し，それらを事前に用意する必要がある．また，強化学習と直接関係ないが，谷[谷96]はリカレントニューラルネットワークを用いて現在のセンサ情報と行動指令から，次のセンサ情報を予測し，移動ロボットの航行実験を行った．但し，ノード数など事前に設計しなければならないパラメータが残っている．

図6 パッサーとシュータの協調行動

Uchibe et al.[UAH97]は，環境のダイナミクスを学習者自身の運動指令を含めて推定する手法を提案している．部分空間法と呼ばれる次数同定の手法を用いており，過去の知覚情報と運動指令の組を入力として，将来の知覚情報を予測し，これによって状態パラメータを推定する．理論的には，観測範囲であらゆる物を無限時間を用いて記述可能であるが，現実には，何らかの規範で同定次数を制限する必要がある．パッサーとシュータが混在する環境でのマルチエージェント学習問題に適用し，実ロボットでの結果を得ている(図6参照)．同時学習は困難なので，最初にパッサー(図右)がある方向にボールを蹴り出す学習を実施後，シューターが転がるボールをシュートする学習を実施する．ともにぶつからないように，障害物回避行動は，事前に学習し，埋め込まれている．現在のセンサ出力の次元を状態空間の次元とした場合の比較として，パスの成功率が約10%から50%以上に，シュートの成功率も同じく約10% から約80%に改善された．彼らは，環境内の個々のエージェント(ゴール，ボール，敵，味方など)のダイナミクスを同定する過程が状態空間構成にあたり，強化学習がエージェントの相互作用を学習する過程とみなしている．実験結果として，物理的に同一の物体(例えば，転がるボール)でも，経験(この場合，タスクの違いによる経験のバイアスが存在)の差異により，推定される状態パラメータが異なったことが挙げられ，ロボットの個性を考える上で興味ある結果と考えられる．今後，協調行動などを実現するとき，このような差異をどのように吸収するかが，課題としてあげられている．

5. おわりに

ロボットの行動獲得手法として，最近注目されている「強化学習」に焦点をあて，実ロボットへの強化学

習適用の問題点として,「状態空間構成」,「より複雑なタスクへの対応」について例を用いて説明した．本文中にも述べたように，示した例は，自由度が少ない移動ロボットであり，行動空間の構成はあまり重要な問題とはなっていないが，今後の方向として，

- マルチエージェント環境での学習を考えると，学習者自身の行為の時間的な分割が重要な問題となる．即ち，自分の行為に対する，相手の行為の同定問題が含まれ，時間軸方向への抽象化により行動空間を構成する問題がクローズアップされれる．
- 我々人間をはじめとする動物は，物理的に多自由度にも関わらず，局面にあわせてキーとなる数少ない制御パラメータで行動しているように見える．このような空間的な行動自由度の抽象化も今後，強化学習に限らず，重要な問題となる．

ロボット学習で対象となっているタスクは，現段階ではまだ，簡単なものであることを否定できない．学習を使わなければ解決困難なタスクをより多く対象とすることで，今後の発展が期待される．

◇ 参 考 文 献 ◇

[ANH96] M. Asada, S. Noda, and K. Hosoda. Action-based sensor space categorization for robot learning. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems 1996 (IROS '96)*, pp. 1502–1509, 1996.

[ANTH96] M. Asada, S. Noda, S. Tawaratumida, and K. Hosoda. Purposive behavior acquisition for a real robot by vision-based reinforcement learning. *Machine Learning*, Vol. 23, pp. 279–303, 1996.

[CM93a] J. H. Connel and S. Mahadevan. "Rapid task learning for real robot". In J. H. Connel and S. Mahadevan, editors, *Robot Learning*, chapter 5. Kluwer Academic Publishers, 1993.

[CM93b] J. H. Connel and S. Mahadevan, editors. *Robot Learning*. Kluwer Academic Publishers, 1993.

[ISI96] H. Ishiguro, R. Sato, and T. Ishida96. Robot oriented state space construction. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems 1996 (IROS96)*, pp. 1496–1501, 1996.

[LCK95] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling. Learning policies for partially observable environment: scaling up. In *Proc. of Conf. on Machine Learning-1994*, pp. 362–370, 1995.

[MA95] A. K. Moore and C. G. Atkeson. Parti-game algorithm for variable resolution reinforcement learning in multidimensional state-sapces. *Machine Learning*, Vol. 21, pp. 199–233, 1995.

[Mah96] Sridhar Mahadevan. Machine learning for robots: A comparison of different paradigms. In *Proceedings of 1996 IROS Workshop on Towards Real Autonomy*, pp. 3–16, 1996.

[McC95] R.A. McCallum. "instance-based utile distinctions for reinforcement learning with hidden state". In *Proc. of the 12th Int. Conf. on Machine Learning*, pp. 387–395, 1995.

[TAH96] Y. Takahashi, M. Asada, and K. Hosoda. Reasonable performance in less learning time by real robot based on incremental state space segmentation. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems 1996 (IROS96)*, pp. 1518–1524, 1996.

[UAH96] E. Uchibe, M. Asada, and K. Hosoda. Behavior coordination for a mobile robot using modular reinforcement learning. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems 1996 (IROS96)*, pp. 1329–1336, 1996.

[UAH97] E. Uchibe, M. Asada, and K. Hosoda. Vision based state space construction for learning mobile robots in multi agent environments. In *Proceedings of 6-th European Workshop on Learning Robots, EWLR-6*, pp. 33–41, 1997.

[WD92] C. J. C. H. Watkins and P. Dayan. "Technical note: Q-learning". *Machine Learning*, Vol. 8, pp. 279–292, 1992.

[Whi91] S. D. Whitehead. "A complexity analysis of cooperative mechanisms in reinforcement learning". In *Proc. AAAI-91*, pp. 607–613, 1991.

[谷96] 谷. 「ロボットにおける認知と自律性の構造: 力学系の見地から」. 日本ロボット学会誌, Vol. 14, No. 4, pp. 4–7, 1996.

[中村96] 中村, 浅田. 運動スケッチ：画像運動情報に基づく単眼視移動ロボット の行動獲得. 人工知能学会誌, Vol. 11, No. 6, pp. 905–915, 1996.

著 者 紹 介

浅田　稔(正会員)

1982年大阪大学大学院基礎工学研究科後期課程修了．同年，大阪大学基礎工学部助手．1989年大阪大学工学部助教授．1995年同教授．1997年大阪大学大学院工学研究科知能・機能創成工学専攻教授となり現在に至る．この間，1986年から1年間米国メリーランド大学客員研究員．知能ロボットの研究に従事．1989年，情報処理学会研究賞，1992年，IEEE/RSJ IROS'92 Best Paper Award受賞．1996年日本ロボット学会論文賞受賞．博士（工学）．日本ロボット学会，電子情報通信学会，情報処理学会，人工知能学会，日本機械学会，計測自動制御学会，システム制御情報学会，IEEE R&A, CS, SMC societies などの会員　〈asada@ams.eng.osaka-u.ac.jp〉