ロボットの行動学習*

浅田 稔†

On Learning Robot Behaviors

Minoru ASADA

Reinforcement learning has recently been receiving increased attention as a method for robot learning with little or no *a priori* knowledge and higher capability of reactive and adaptive behaviors. This paper presents a framework of the reinforcement learning, and several issues in applying the method to real robot tasks. Then, examples of real robot applications, especially vision-based reinforcement learning methods are introduced to show how they cope with these issues.

Key Words: Robot Behavior, Reinforcement Learning, State Space Construction, Action Space, Behavior Coordination, Hidden States

1 はじめに

AIとロボティクスの究極の目標は，動的な環境との相互作用を通じて自律的に作業を遂行するロボットを実現することであり，これまで，多くの研究者が熟考的かつ段階的アプローチでこの問題に取り組んできた．しかしながら，システムが複雑化するに連れて，環境の変動に対処できない可能性があり，これらの拡張だけでは自律的なエージェント(ロボット)を実現困難であることが指摘されている[1]．この問題に対処するために，Brooks[2]は，行動規範型ロボットを提唱し，彼の研究グループは，いくつかの行動規範型ロボットを作成した[3]．これらのロボットは，環境に対して反射的行動をとることができるが，ある目的地に辿りつくような目的行動を生成する能力にかけること，個々の行動モジュールは従来型のコード化作業が必要であることなどが欠点として挙げられている．

これに対し，最近，反射的かつ適応的な行動を獲得できるロボットの学習法として，強化学習が注目を浴びている[4]．この学習法の最大の特徴は，環境やロボット自身に関する先験的知識をほとんど必要としないところにある．強化学習の基本的な枠組みでは，ロボットと環境はそれぞれ，離散化された時間系列過程で同期した有限状態オートマトンとしてモデル化される．ロボットは，現在の環境の状態を感知し，一つの行動を実行する．状態と行動によって，環境は新しい状態に遷移し，それに応じて報酬をロボットに渡す．これらの相互作用を通して，ロボットは与えられたタスクを遂行する目的行動を学習する．

本稿では，まず強化学習の基本的な枠組を述べた後，実際のロボットへ適用する際の問題点として，状態空間の構成法，多重タスクへの適用を指摘する．そして，実際の適用例として，視覚誘導型移動ロボットの強化学習を示し，現状の問題点とその対処法，今後の展望などを述べる．

2 強化学習の枠組

強化というと，アメリカの行動心理学者スキナーのスキナーボックスが思い出される．鼠を箱のなかにいれ，そのなか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そうである(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)状態遷移する．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)

*原稿受付　平成8年8月9日
†正員，大阪大学工学部 〒565 吹田市山田丘2-1

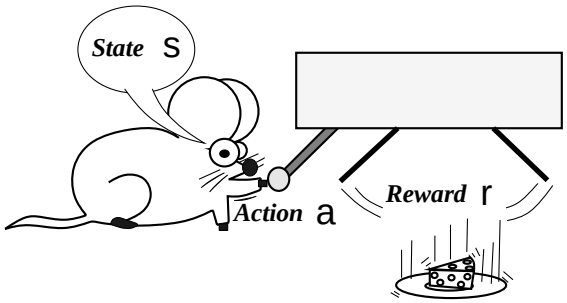


Figure 1: Skinner's box

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**[5]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，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．

3　ロボット学習の課題

強化学習の役割は自律的なエージェントを実現する上で非常に重要であるが，その意義は，より大きく複雑な問題にどの程度適用可能かに依存する．強化学習を始めとするロボット学習を実際のロボットに適用する際の三つの基本的な問題点を以下にまとめる．

- **状態・行動空間の構成：**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い．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．例えば，2次元格子状の世界で，ロボットの行動は格子上の上下左右への移動のいずれかであり，状態として格子の座標を対応させるものである[6]．このような状態空間の構成法は，実際のロボットシステムとコンピュータシミュレーションとのギャップを広めている．それぞれの空間は，ロボットが実際感知したり行動できる物理世界と対応すべきと考えられる．しかも，学習が正しく収束できるように構成されねばならない．更に，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を決定する問題も考慮されねばならない．

- **報酬関数の構成及び学習の高速化・効率化(発達と導き)：**Q学習の更新式(1)では，状態遷移毎に即座の報酬があたえることが出来るが，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が，長い学習時間を必要とする．これを解決する学習の高速化の手法としては，事前にサブタスクに分解する手法[7](分解のための知識を事前に必要とする)，外部から評価者による学習[8](随時，行動の是非を判断する厳密な知識(神様)が必要)，他の学習者の結果を共有する方法[8](経験を共有することによる学習の並列実行による学習時間の短縮化)，簡単なタスクからの実行[9](厳密にやさしさ順で学習を実行すれば，指数オーダーから線形オーダーに短縮．問題は，やさしさの順をどれくらい正確にしっているか?)などが提案されているが，それぞれに一長一短と考えられる．

以下では，実際のロボットに適用された例をもとに，これらの問題点を具体的に明らかにする．

4　状態・行動空間の構成問題

通常の強化学習では，学習が収束するように上手に離散化された状態と行動が定義されているばあいが多い．実ロボットの例では，Maes and Brooks[10]が6本脚ロボットの歩行実験で，状態として脚及び腹部の着地(接触センサーのON/OFF)をもとに，歩行を学習させているが，状態数が少なく，反射的で簡単な実験例であり，状態・行動空間の構成が重要な問題にならない．しかしながら一般には，学習が収束可能な状態・行動空間を構成することは容易ではない．

ロボットがタスクを遂行するために必要十分な情報を含む状態空間の構成は，ロボットの行動能力に依存する．また行動空間もロボット自身の知覚能力に依存

し，相互に規定しあう．この問題に対し，行動空間を先に固定して，状態空間を構成する手法が提案されている．Chapman and Kaelbring[11] は，TV ゲームの主人公が敵と戦って目的を達成するタスク設定で,「敵を撃つ」,「障害物を回避する」などの構造化された行動をもとに,「敵が部屋にいる」,「ドアが開いている」などの既に抽象化された状態の真偽 (オン/オフ) をビット列とする入力ベクトルを，報酬をもとに分割する手法を提案している．しかしながら，もとの状態が既に抽象化されており，一般的なセンサの実数値の連続空間を対象とする問題には適用できない．

Dubrawski and Reingnier [12] や Kröse and Dam[13] らは，移動ロボットの障害物回避のためのソナー情報の抽象化手法を提案しているが，障害物回避などの反射的なタスクを想定しているので，行動の物理的単位が固定されていても問題は生じにくいが，視覚情報などを利用して遠くにある目標物に到達するタスクなどを想定した場合，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 [9] では，状態数を低減する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，同じ行動をとっても状態遷移のバラツキが大きくなり，学習が進まない．この問題は**「状態と行動のずれ問題」**と呼ばれ，これに対し，状態が変わるまで，同じ動作を続け，その一連の動作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が，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．

そこで，野田ら [14] は，見掛けが異なっても，ボールとゴールが正面にあれば，直進さえすれば，ゴールにたどり着けるので，ゴールへの行動が同じであれば，同じ状態としてまとめられないかと考え，以下の手法をとった．まず物理的に定められた単位時間に発生されるモーターコマンドの結果として生じるロボットの動作を「行動要素」と定義し，この行動要素に基づいて，ゴール状態や既に獲得された状態に，一種類の行動要素の可変時系列で到達できる入力ベクトルの集合を状態と定義，また「状態遷移を引き起こす同一の行動要素の時系列」を一つの行動とした．そして，初期状態としてゴール状態とその他の二つにしか分かれていなかった状態空間を再帰的に分節化していった．Figure 2と表1に比較結果を示す．図では，ボールの大きさとゴールの大きさの 2 次元に投影している．格子状の箱は，先の研究で人間が与えた状態分割で，楕円状に投影されている領域が自律的に分節した結果である．形状が大きく異なること，また大きさも異なり，探索時間が激減し，シュートの成功率も上がっている．

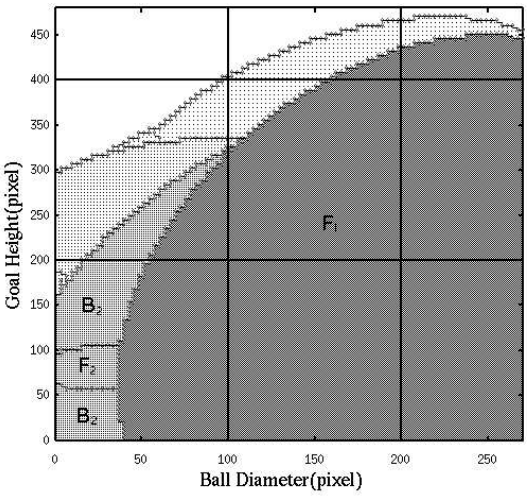


Figure 2: State space segmentation

Table 1: Result

	# of states	search time (1/30sec.)	success rate (%)
programmer	243	500M*	77.4
proposed method	33	41M	83.3

* indicates learning time.

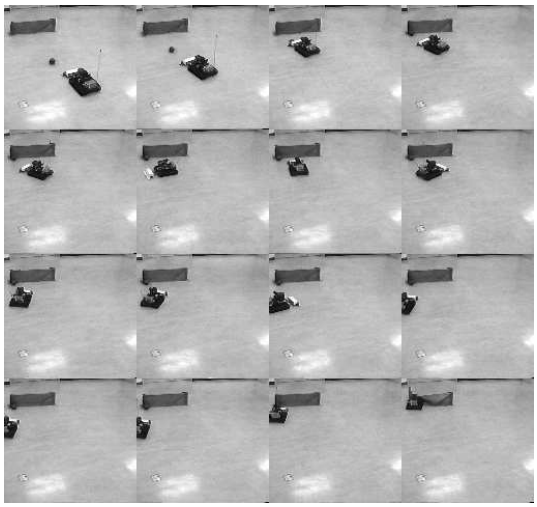
実ロボットがシュートする様子の一例を図3に示す (左上から横方向へ)．約 1 秒毎の動きを示している．(a) に，シュートの様子を，(b) にその時ロボットから見た環境の様子を示す．この例では，最初シュートを試みたが，ゴールの左隅にボールが留まり，ゴールできていないので，左右に振りながらシュートできる地点まで後退し，再シュートして成功している．

この例では，以下の二つが重要と考えている．

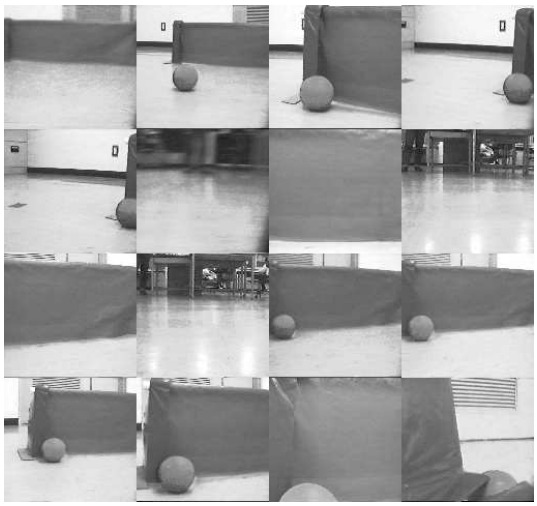
1. 人間が与えた状態空間はロボットにとって最適とは限らず，ロボットが自らの経験から自分が見切るべき状態を決定すべきで，世界の分節化問題と同じであると考える．すなわち，主体と環境に依存して，世界の見方が変わる．

2. 本来，状態と行動空間の構成問題は,「鶏と卵」問題に類似し，相互に密接に関連している．何らかの最低限の拘束を導入しないと，相互に規定できない．ここでは，行動の最小単位を規定し，状態

が変化するまで，同じ「行動要素」を実行し続けることにより，続ける長さをパラメータとして行動空間と状態空間を同時に構成していると考えることも可能である．



(a) shooting behavior



(b) a sequence of images taken by the robot

Figure 3: Shooting behavior by the learning method

後者は,「行動要素」を拘束として，行動空間を時間軸に抽象化したと考えられるが，物理的な自由度が2で，行動空間の空間的な抽象化の対象としては自由度が少なく，その結果，抽象化が知覚中心となっていると考えられる．

上の例では，画像情報からボールとゴールの領域特徴を抽出し，それらを状態空間の軸としているが，一般の画像を対象とした場合には，何を特徴とすればよいか明確ではない．そこで，画像の画素単位の情報をもとに状態空間を構成する手法も提案されている．佐藤ら [15] は，移動ロボットから得られる視覚情報を多変量解析し，障害物回避のための状態構成問題を扱っている．また，人間の教示データを効率よく表現するために，画像情報と運動指令の非線形相関を表現する手法も提案されている [16]．膨大な画像情報から効率的に，行動との相関をとるために，佐藤らの方法でも，人間側が教示データを与えているが，ロボットが人間と同じ構造をもっているとは限らないので，教示データに誤りがあったり，最適でない場合にでも，修正したり汎化できる能力が必要であろう．

5　学習の効率化 - 複雑化への対応 -

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 [8] 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，それらを統合することである．Mahadevan and Connel[7] 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つに前持って分割し，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．但し，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．また，バンパーセンサー，ソナーなどの近接センサのみを利用しているので，作業の遂行が局所的であり,「箱を指定された場所に運ぶ」などの大局的な目的行動を獲得することには向いていない．

内部ら [17] 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 [9] と，別に獲得した回避行動を統合することを考えたが，以下の問題が含まれていた．

- 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，知覚見せかけ問題 [18] による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．

そこで，彼らは前者の問題に対し，最尤推定を用いて，隠れ状態を推定・識別した．また後者に対しては，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た．表2にシミュレーション結果を示す．比較のために，回避行動を伴わな

いシュート行動のみも考慮した．表から分かるように，再学習した結果が，シュート率，回避率，シュートステップ数で最良であった．実ロボットに適用した様子を図4に示す．最初ボールを探すために後退し(第2画面)，右前進してボールを捉え(第4画面)，ゴールキーパーを避けてシュートしている様子が窺える．

Table 2: Comparison with various methods

method	success rate (%)	mean steps between collisions	mean steps to the goal
only shooting	46.7	43.1	286.9
simple sum	33.2	77.5	231.2
switching	39.2	98.0	414.4
learning	46.7	238.1	128.3

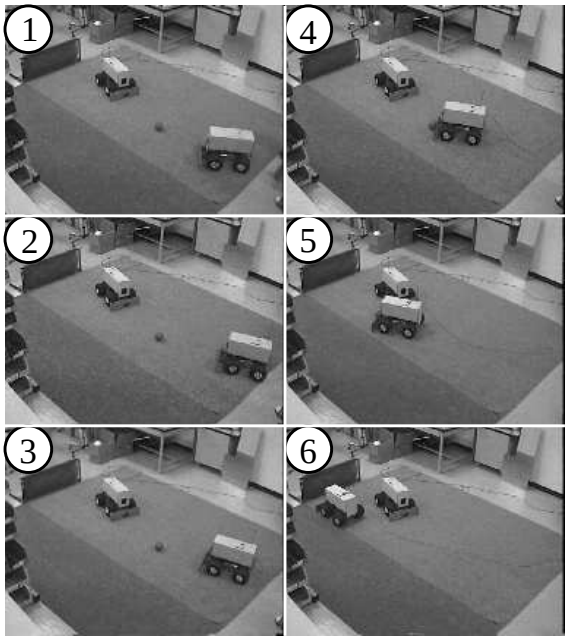


Figure 4: Shooting behavior avoiding a goal-keeper

個別に学習された行動を統合する手法として，それぞれの行動価値関数の性質を利用する方法も考えられる．Nakamura and Asada[19] は，画像上の見掛けの速度分布(オプティカルフローと呼ばれている)を実時間で検出できる画像プロセッサーを用いて，フローパターンと到達行動や回避行動のマッピングを個別に強化学習し，回避行動の行動価値関数のばらつきから，回避行動への切替え条件を求めている．

更に，彼らは[20]は，障害物を回避して目標物に到達するタスクで，それぞれのサブタスクに分解する事なく，単一の枠組の強化学習で行動を獲得した．センサー情報として先のフローによる運動情報とステレオ視から得られる視差情報(距離の逆数に比例)を用いて，目標物の状態記述を工夫することで，障害物を陽に記述することなく状態空間を構成し，結果として,「観測のための行動」や「行動のための観測」を区別する事なく，目的行動が達成されている．

より複雑なタスクとして，マルチロボットによる集団行動の学習がある．一般には，複数のロボットが共に学習中であると，相手のロボットの行動戦略が確定していないため，行動学習が困難である事が指摘されている．そこで，Mataric[21]は，独立で干渉しない個々の協調的活動(譲り合いや助け)を模倣することで報酬を与え，グループとしての社会的な集団行動を学習させている．

6　発達と導き

これまでは主に，ロボット内部の学習アルゴリズムを中心に述べてきたが，目的の行動をロボットに学習させるための環境設定も重要な問題である．サッカーロボットの例でも，相手がいない環境[9]ですら，ボールをゴールにシュートさせるために，ボールをゴールの近くに設置しシュートしやすい環境から始めた．またゴールキーパーがいる場合でも，ゴールキーパーが最初から上手だと，プレーヤーは全くシュート出来ず，自信を持てない．そこで，やさしいタスクから学習する規範(Learning from Easy Mission: LEM)[22]に則り，最初ゴールキーパーを静止させ，徐々に速度を挙げることで，最初から同じ速度で動くキーパーを相手にした場合と比較して格段の学習時間の短縮が可能になった[23]．Figure 5に，その効果を示そう．最初にゴールキーパを静止させ，次に学習ロボットの最大速度の半分(左の矢印の時刻)，最後に同じ速度まであげた場合(右の矢印の時刻)のシュートの成功率が実線，最初から同じ速度を持たせた場合が破線で示されている．

人間を始めとする動物でも，様々な教示を通じて物事を学ぶ[24, 25]ことから，ロボットにおいても，直接指示の教示[26]が重要な役割を果たす事は否めない．但し，ロボットにとって教示された内容が解釈可能であり，汎化可能な枠組そのものをロボットの能力として埋め込むことも容易ではなく，自らの学習能力と併せて発達させる必要がある．

7　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

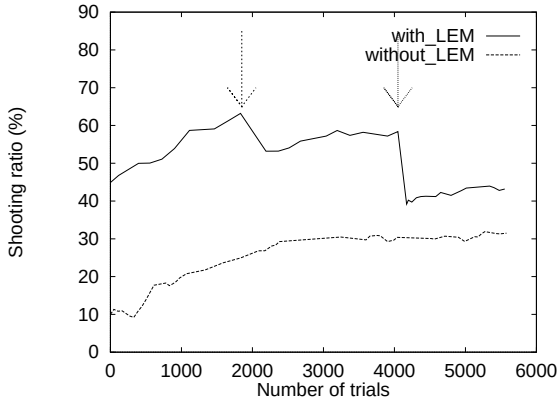


Figure 5: Efficient learnig from easier tasks

習適用の問題点として,「状態空間構成問題」と「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．しかし，多自由度のロボットを制御する場合には，状態空間の構成問題と同様，行動空間の構成問題が重要になり，相互に規定しあう困難な問題となり，これを解決する手法が望まれる．

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．

References

[1] R. A. Brooks. “Elephants don’t play chess”. In P. Maes, editor, *Designing Autonomous Agents*, pp. 3–15. MIT/Elsevier, 1991.

[2] R. A. Brooks. “A robust layered control system for a mobile robot”. *IEEE J. Robotics and Automation*, Vol. RA-2, pp. 14–23, 1986.

[3] P. Maes. “The dynamics of action selection”. In *Proc. of IJCAI-89*, pp. 991–997, 1989.

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

[5] C. J. C. H. Watkins. *Learning from delayed rewards”*. PhD thesis, King’s College, University of Cambridge, May 1989.

[6] S. Whitehead, J. Karlsson, and J. Tenenberg. “Learning multiple goal behavior via task decomposition and dynamic policy merging”. In J. H. Connel and S. Mahadevan, editors, *Robot Learning*, chapter 3. Kluwer Academic Publishers, 1993.

[7] 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.

[8] S. D. Whitehead. “A Complexity Analysis of Cooperative Mechanisms in Reinforcement Learning”. In *Proc. AAAI-91*, pp. 607–613, 1991.

[9] 浅田, 野田, 俵積田, 細田. “視覚に基づく強化学習によるロボットの行動獲得”. 日本ロボット学会誌, Vol. 13:1, pp. 68–74, 1995.

[10] P. Maes and R. A. Brooks. “Learning to coordinate behaviors”. In *Proc. of AAAI-90*, pp. 796–802, 1990.

[11] D. Chapman and L. P. Kaelbling. “Input Generalization in Delayed Reinforcement Learning: An Alogorithm and Performance Comparisons”. In *Proc. of IJCAI-91*, pp. 726–731, 1991.

[12] A. Dubrawski and P. Reingnier. Learning to Categorize Perceptual Space of a Mobile Robot Using Fuzzy-ART Neural Network. In *Proc. of IEEE/RSJ/GI International Conference on Intelligent Robots and Systems 1994 (IROS ’94)*, pp. 1272–1277, 1994.

[13] B.J.A. Kröse and J.W.M. Dam. Adaptive state space quantisation for reinforcement learning of collision-free navigation. In *Proc. of IEEE/RSJInternational Conference on Intelligent Robots and Systems 1992 (IROS ’92)*, pp. 1327–1332, 1992.

[14] 野田, 浅田, 細田. 強化学習によるロボットの行動獲得のための状態空間の自律的構成. 第5回ロボットシンポジウム予稿集, pp. 145–150, 1995.

[15] R. Sato, H. Ishiguro, and T. Ishida. State-space construction considering robot properties. 第13回日本ロボット学会学術講演会予稿集, pp. 453–454, 1995.

[16] G. Z. Grudic and P. D. Lawrence. Human-to-Robot Skill Transfer Using the SPORE Approximation. In *Proc. of IEEE Int. Conf. on Robotics and Automation*, pp. 2962–2967, 1996.

[17] 内部, 浅田, 野田, 細田. 視覚に基づく強化学習による移動ロボットの多重タスク遂行のための協調行動の獲得. 第21回 人工知能基礎論研究会 (SIG-FAI-9403), pp. 25–32, 1995.

[18] S. D. Whitehead and D. H. Ballard. “Active Perception and Reinforcement Learning”. In *Proc. of Workshop on Machine Learning-1990*, pp. 179–188, 1990.

[19] T. Nakamura and M. Asada. Motion Sketch: Acquisition of Visual Motion Guided Behaviors. In *Proc. of IJCAI-95*, pp. 126–132, 1995.

[20] T. Nakamura and M. Asada. Stereo Sketch: Stereo Vision-Based Target Reaching Behavior Acquisition with Occlusion Detection and Avoidance. In *Proc. of IEEE Int. Conf. on Robotics and Automation*, pp. 1314–1319, 1996.

[21] M. J. Mataric. Learning to Behave Socially. In *Proc. of the 3rd Int. Conf. on Simulation and Adaptive Behaviors – From animals to animats 3 –*, pp. 453–462, 1994.

[22] M. Asada, S. Noda, S. Tawaratsumida, and K. Hosoda. Vision-Based Reinforcement Learning for Purposive Behavior Acquisition. In *Proc. of IEEE Int. Conf. on Robotics and Automation*, pp. 146–153, 1995.

[23] M. Asada, E. Uchibe, and K. Hosoda. Agents That Learn from Other Competitive Agents. In *Proc. of Machine Learning Conferen Workshop on Agents That Learn from Other Agents*, pp. 1–7, 1995.

[24] リーキー著, 馬場悠男訳. 「ヒトはいつから人間になったか」. 草思社, 1996.

[25] 正高. 「身体運動は言語獲得にどのような役割 を果たすか」. 日本ロボット学会誌, Vol. 14, No. 4, pp. 31–34, 1996.

[26] 國吉. 「実世界エージェントにおける注意と視点-情報の分節・統合・共有-」. 人工知能学会誌, Vol. 10, No. 4, pp. 507–514, 1995.