特集　ロボカップ　　３．ロボットプレーヤの感覚と学習

浅　田　稔

実際のロボットを使ってサッカープレーヤーを作ることに，どれくらい意味があるかなんて，最初からそんなに真剣に考えていた訳ではない．本稿の次の解説の著者の稲葉さんたちが，ラジコンのサーボを使って，面白いロボットを作っているのを横目で眺めていて，簡単でいいから何か応用がないかなぁと思っていたこと．もう一つは，ロボットで新鮮味のあるテーマで何か面白そうなことなかなぁと考えてていたころ，強化学習関連の本[1]の輪講に加わり,「こりゃおもしろそうだ」と思ってしまったことがきっかけである．これまでの研究で，我々が学んだことは，至極当り前だけど，以下の二つである．**「ロボットは最初から難しいことはできない．簡単なタスクからやらせよう.」**そして，**「ロボット自身の能力で世界をみないといけない.」**本稿では，これまでロボットプレーヤの感覚と学習の関連で我々のグループで行って来た研究例を裏話を含めて紹介する．

1　は　じ　め　に

鉄腕アトムやR2D2，C3PO，T2やT1000など，漫画やSF映画に登場するロボットたちは，どれも(誰も?)イキイキしており，環境や状況の変化を一瞬の内に察知し，時には失敗するものの，うまく仕事をこなしているように見える．このようなロボットを実現することは，AIとロボティクスの究極の(ということは，当面実現できそうにない)目標である．では，現在のAIやロボット研究者は，何を研究すべきか? 言語理解が大切だ，運動能力も必要，ちゃんと世界を見て欲しい，学習能力が不可欠だ，ええ，もちろん，これらの研究すべて大事だけど，これって別々にやっててどれくらい意味があるのだろう．もちろん，全て同時に研究はできないけど，少なくとも，見て-判断して-動いてのサイクルが充分短くないと，とてもイキイキとは写らない．しかも，前に経験したことをちゃんと活かして欲しい．では，タスクを簡単にして，これらのことが実現できないか，と考えれば,「サッカーロボット」は,それなりに，いいテーマに見えて来る．ここらあたりに関しては，本稿より前の解説で充分説明がなされているだろうから，深入りしないでおこう．

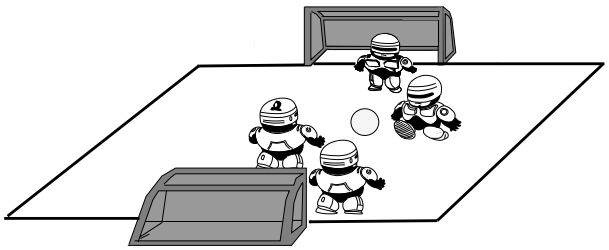


図 1: **ロボカ(コ)ップ?**

1993年の秋，当時，O大のT先生(その時，学部長だったかなぁ? 現在，W大学)が，AIやロボットの身内の研究者を誘って，強化学習関連の本[1]の輪講会を開いた．そこで論文を紹介したのが，強化学習とのつき合いの始まりである．強化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で，この特徴から，ロボットの学習法として，非常に魅力的に見えてくる．但し，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．ところが，学習屋さんの興味は，実際のロボットを使って検証することではなく，学習の収束証明，高速化のための理論的考察などにある[2]．ほとんどがコンピュータシミュレーションによる結果で，多くは，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[3]．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．なぜなら，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からだ．特に，視覚を用いた実ロボットの強化学習応用は，当時我々の知る限り見当たらなかった．

そこで，我々は，強化学習が実ロボットで，ほんとうにどれくらい使えるのかを試す上で，サッカーロボッ

トを選んでみた．以下では，まず強化学習について少し詳しい説明をするけど我慢してつき合って欲しい．最初はボールをゴールにシュートするだけのロボットの学習だけど，我々の最初の成果なのでちょっと詳しく説明する．人間が事前に状態空間を分割した場合[4]とロボットが経験に基づいて状態空間を分割し，それを使って行動を決定した場合[5]がある．次に，一人でシュートだけするのは寂しいので，ゴールキーパーを設定し，ゴールキーパーを避けてシュートする行動について学習により獲得するもの[6]，ゴールキーパーの行動をスケジュールする事により学習者の効率を上げたもの[7]を紹介する．

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を思い出される読者も多かろう．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．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)が与えられるという仕組みである．強化学習は，これを確率的動的計画法(Stochastic DP)の枠組で定式化したものである．

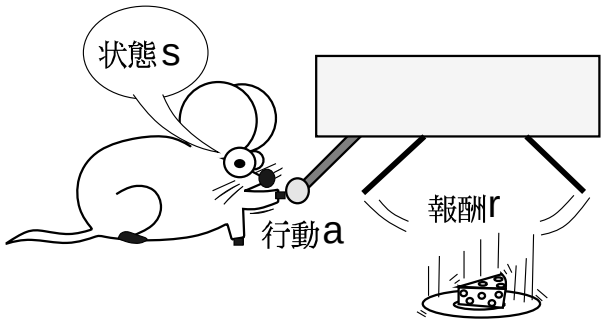


図 2: **スキナーボックス**

鼠は箱の中ので，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DP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**[8]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　最初はシュート行動だけ

最初から敵のチームを相手に味方と一緒に戦おう何て，とても無理なので，第一ステップとして，とにかくボールをゴールにシュートさせる行動を学習させた．**図3**に示すように，環境内にはボールとゴールしかなく，ロボットに入ってくる情報は，それらの画像情報のみである．最初にコンピュータシミュレーションで学習させ，そのあと実機に学習結果を写して，実行させた．実ロボットはブルトーザーのラジコンを改造したものである．**強化学習を適用するには，状態空間を定義しておかないといけない．**得られる情報は，ボールとゴールの画像情報だけなので，それらの特徴を反映した状態として，ボールの位置と大きさ，ゴールの位置と大きさ，及び向きについて**図4**に示す状態空間を用いた．

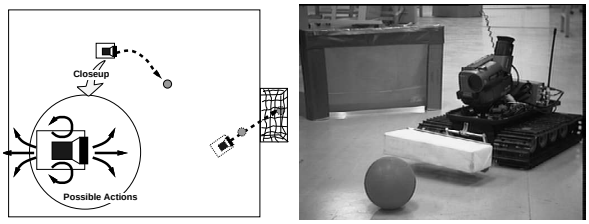


図 3: **タスクとロボット**

行動の方は，ブルトーザの二つの独立したモータへの前進，停止，後退命令の組合せで**図3**に示すように，ストップを除いた8種類の行動からなる．尚，これらの物理的意味などは，ロボットが知る由もない(ことにしている)．これらの状態と行動で簡単に学習できれ

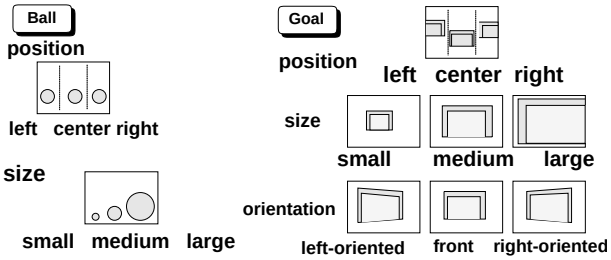


図 4: **ボールとゴールの状態空間**

ば，実ロボットでの検証なんて必要ない．うまく行かないから面白い．最初の問題は，**実際のセンサーやアクチュエータを反映した状態と行動は，必ずしも同期しない**ことである．**図5**に示すように，ゴールから遠く離れた所にいる場合，前進行動を一回とっても状態は変わらない．ところが，中間位との堺目だと1回で，状態が変わってしまう．つまり，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．そこで，ここでは，**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した．**これは，以下で述べるその他の例でも，基本原理として用いている．

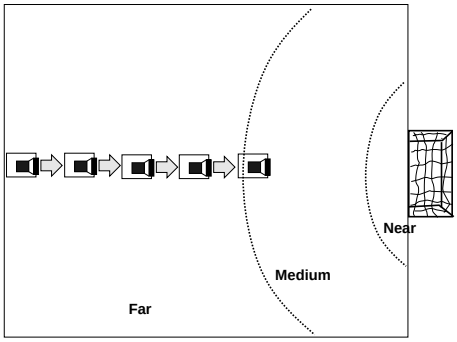
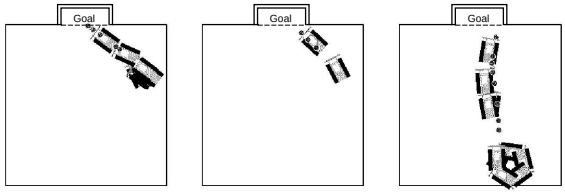


図 5: **状態と行動のズレ問題**

以上のように，状態と行動を設定して学習を始めたが，なかなか収束しない．いつまでたっても馬鹿のままである．教育方針は，全くの放任で，設定としては，行動価値関数の初期値ゼロ，報酬はボールをゴールにシュートしたときのみで，それ以外はゼロである．ゴールした時しか報酬がもらえないので，やみくもにトライしてもほんとんどだめだからと考え，ゴール状態との距離に応じて報酬を適当に与えたら，極大値がいっぱいできて，これまた全然収束しなかった．適当でなく，きっちり報酬を決めたら収束するのだろけど，それなら，最初から制御則を書くのと同じで，学習の意味がない．考えてみれば，これは当り前で，小さな子どもに始めてサッカーをやらすようなもので，どこにボールがあるのやら，どこへ蹴ったらいいのやら，はてはどうすれば蹴ったことになるのかわからないないままに，試行させているようなものである．そこで，How は一切教えずに，自信を与えるために，最初は簡単な状況から学習を始めさせた．すなわち，ボールとロボットをゴールの近くに設定し，偶然成功する確率を高め，そのあたりでの学習が収束すれば，ちょっとゴールから離れたところに設定する．これで，やっとうまく学習が進み，シュートに成功し始めた．全く，子ども (大学生，院生，教官も?) の教育とほとんど同じである．



(a) $\gamma = 0.999$　　(b) $\gamma = 0.6$　　(c) 一連の行動

図 6: **シミュレーション中の行動**

シミュレーション中のロボットの行動例を示そう．**式 (1)** なかのγによって，表出するロボットの行動が変わってくる．**Fig.6(a)** では，γの値が大きいので，どんなに時間を掛けてもゴール時の報酬が，あまり変わらない．そのため少しでも確実にシュートしようと，よりよい位置に移動してからシュートしている (保守系のおっとり型) のに対し，(b) では，γの値が小さく，早くゴールしないと報酬がもらえないので即座にシュートしている (革新系の速攻型)．尚，(a,b) では，比較のためにスタート地点は同じである．(c) では，学習した政策を用いた一連の行動を表す例を示した．最初にボールを見失い，発見し，ドリブルして，最後にシュートしている．シミュレーションはうまくいっている．

実際のロボットの制御は，稲葉さんたちの Remote-Brain の手法 (これの次の解説参照) を利用している．状態識別を実時間 (1/30 秒) で終えなければばらないので，ボールを赤に，ゴールを青にして処理速度を上げている．処理結果の一部を**図7**に示す．シミュレーションとの違いは，ビデオノイズが多い事，キャタピラが滑ること，ボールが思わぬ方向によく転ぶことであ

るが，状態空間を粗くしているため，これらの影響は少ない．すなわち，多少の画像処理の精度に影響されない．キャタピラが滑べるのは動作指令を変える時であるが，状態が変化するまで，同じ動作命令を実行し続けるので，あまり起こらない．ボールの転びは，厳密にはマルコフ性に反しているのだが，割と粗い状態空間を構成しているので，あまり強く影響がでていない．但し，粗いゆえにゴール直前での微調整が効かないことも確かで，シュート率は，驚くほど良くはない．

実機での実験結果はカラー写真のページを参照．約1秒ごとのロボットの動きと，ロボットからみた画像を16枚示している．一回左隅へトライし，失敗したので一旦腰を振りながら後退し(このときボールを見失っている)，左側からゴールへシュートしている．ゴールはロボットを傷つけない様に簡単に壊れるようにしている．なぜなら，一回毎にロボットが潰れてはこちらの予算と労力が持たない．大事にしているのである(ロボット屋の常識!)．

図 7: **画像処理結果**

4　もう一つのシュート行動

先の例に関して,「既存の制御手法使えばいいじゃない」,「ファジー制御簡単にできるわ」というひとが少なからずいて，比較した[9]．まず，環境やロボットの構造パラメータを与えてボールが画像の真中に来るようなフィードバックをかけて実施してみたが，若干の工夫を必要とした．シュート率は，改善されたが，ゴールまでの平均ステップ数は，学習によるものより多かった．また，ファジールールは，先の学習プログラムを書いた学生が，学習中のロボットの挙動を学習してさらっとファジールールを書いた．そのシュート率は，確かに若干よくなったが，劇的に変わる物ではなかった．注意したいのは，それらは，人間の介在が大きいものである．というと，こんどは，ボールやゴールの状態空間を与えた時点で教えたも同然だとおっしゃる方がおられる．じゃー，状態空間をロボットが作ればいいでしょう，ということで，もう一つのシューティング行動の学習につき合って頂きたい．

状態空間を構成する問題は，基本的で重要で非常に困難な問題である．この例では，すでにボール(赤領域)，ゴール(青領域)が抽出されているから，かなりやさしくなっているかもしれない．それらの領域特徴を状態空間の次元として採用したが，一応，これも自動抽出を試みた．多数の動作から得られる画像情報から面積，重心，モーメントなどの基本特徴を並べ，主成分解析で主成分を求めると，ほぼ自明だが，ボールやゴールの位置，大きさ，傾きなどに相当する特徴が対応した．次の問題は，この軸をどのように分けるかの「分節化」の問題である．先に述べたように，基本原理(「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した.」)から，逆に**ゴール状態に到達するまでの動作が変化しないならば，それらの見掛けが異なっていても，同じ状態に区分する**方針をとり，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．**図8**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．

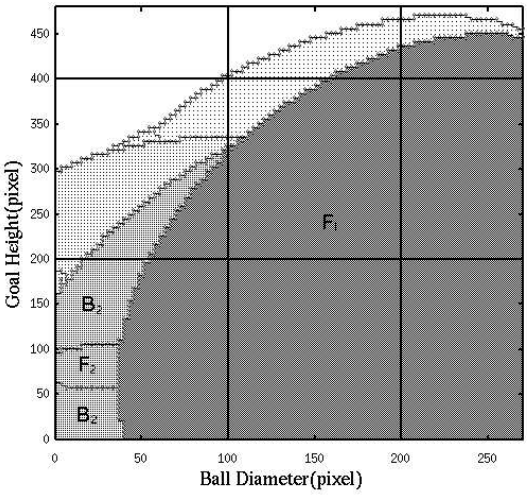


図 8: **状態の自律的分節化**

5　ゴールキーパが相手

たった一人でシュート行動だけしていると，あまりに寂しいので，ゴールキーパを立たせ，それを避けて

表 1: **人間が与えた状態空間と自律的に構成された空間**

	状態数	探索時間 (1/30 秒)	成功率 (%)
人間が与えた	243	500M*	77.4
自律的に構成	33	41M	83.3

* Q学習の時間を指す.

シュートする行動の学習を行った．ここでの問題は，まずシュート行動の獲得だけでも状態空間が結構大きいのに，ゴールキーパを含めた状態空間で最初から学習するのは，不可能に近いことだ．因みにQ学習の収束時間は状態空間のサイズの指数オーダーである．そこで，既に獲得されてるシュート行動を活かす方法を考えた．すなわち，ゴールキーパを避ける行動を別途学習し，それと統合させることである．ここで，次の問題が生じた．どうやって統合するか? すなわち行動の切替え条件をどう設定するかである．サッカーゲームの場合，ゴールの近くであれば，ぎりぎりまで相手に詰め寄るが，そうでなければ，イエローカードがあまり欲しくないので，危険なことはしたくない．つまり，状況に依存するわけだ．それを全て人間が調べる訳にもいかないので，学習で切り分けを実現した．

まず，ゴールキーパの回避行動の学習である．回避行動は，具体的な行動指令は唯一ではなく，衝突以外の行動を選択できればよい．そこで，衝突行動をシミュレーションで学習させ，その行動価値関数の符号を変えて用いればよい．ここで，気をつけなければいけないのは，**式 (1)** 中のγの値である．当初，シュート行動と同じ程度の0.8に設定したが，臆病者で，ゴールキーパをみるといつまでも後退する．困ったもんだ．これは，行動価値があとあとまで響いているためで，小さな値に設定することにより，反射的な行動が獲得できる．もちろん，回避行動の明示的なプログラミングそのものはそんなに困難でないが，シュート行動との統合がやりやすいように，強化学習で獲得させた．

なお，シュート行動の状態空間は**図 4**と同じで，回避行動の場合は，ボールをゴールキーパに変えた場合と同じものを使った．考えられる統合法として，まず二つの行動価値関数の単純和を考えた．それぞれの行動は他方の行動に関する状態を考慮しないので，当然の事ながら，極大値にはまったり (停留点問題)，意味のない行動をとることがある．次に，切替えを考えた．ある条件に従って切替えるが，その条件を決定することは難しい．

単純和や切替えによる行動の統合問題を解決するために，もう一度学習を実施した．初期値として単純和による行動価値関数を設定し,「ボールがゴールキーパに隠される」などのあらたな状態を付加した．そして，それらの近傍では，未経験領域なのでランダムに探索し，それ以外では，既存の行動価値の高いものを選択することで学習の高速化をはかった．**図 9**と**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．図には，ボールが隠されている状況からシュートしている様子を示している．

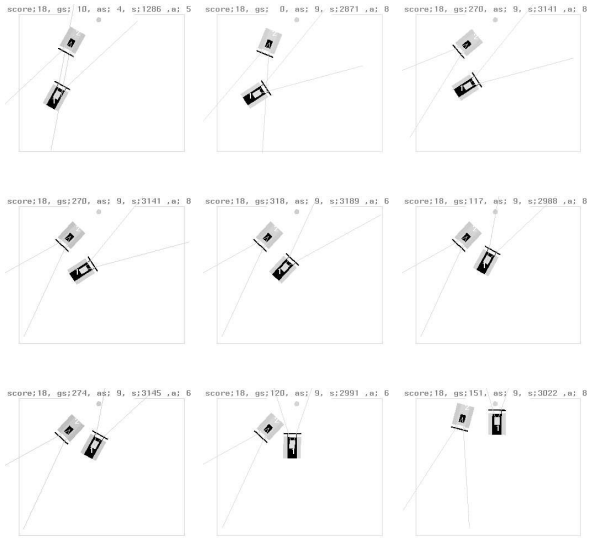


図 9: **ゴールキーパを避けてシュート**

表 2: Simulation result

統合法	成功率 (%)	平均衝突間ステップ数	シュートまでの平均ステップ数
シュート行動のみ	46.7	43.1	286.9
単純和	33.2	77.5	231.2
切替え	39.2	98.0	414.4
学習	46.7	238.1	128.3

これで，ゴールキーパがいても，大丈夫と思うのは早合点．上手なゴールキーパだったらどうする．実は，上の例では，わざとゴールキーパの能力を学習ロボットより少し劣らしている．もし，上手なゴールキーパだったっら，一度もシュートできず，自信喪失となって

しまう．逆に，下手なゴールキーパで自信を持たせて，少しずつ慣らせれば，収束が早い．**図10**のその効果を示そう．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，最後に同じ速度まであげた場合のシュートの成功率が実線(LEMは，Learning from Easy Missionsの略)，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．実機での実験の様子はカラー写真その2を参照．最近購入した4駆のラジコンを改造し，筆者の趣味で黄色のユニフォーム(黒との縞模様が理想だが，工事現場と間違われるのと画像処理がややこしいので)を着せた．まだまだ動きが鈍いが，2台がなんとか動いている様子がお分かりと思う．

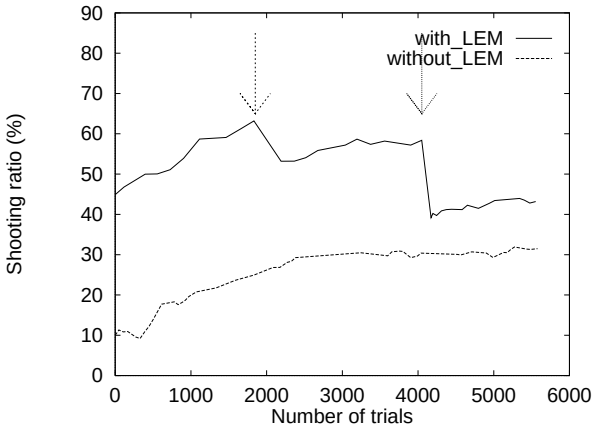


図 10: **やさしいタスクからの学習による効率化**

6 お わ り に

まだまだ，問題が山積みである．マルチエージェントによる協調，競合，競技の問題はデフォルトとして，それらを実現するハードウェアの整備が大変である．例えば，ボールやゴールだけでなく，フィールドラインや，コーナーポール，敵，味方を識別するとなると，現在の簡単なカラー処理だけでは，困難であろう．処理アルゴリズムはいくらでも考えられるが，実時間処理を考えると，何をサボればいいかを考える必要がある．研究テーマとしては，ロボットが自律的にこれを発見して欲しい．いずれにしろ，複数のロボットが互いに敵，味方に分かれ，動きが多少不細工でも，実環境でゲームをしてくれれば，なかなか壮観であろう．

最後に，謝辞．文部省の科研補助金ありがとうございました．また，研究の実施にあたり細田耕博士，卒業生の野田彰一君(現，日立製作所)，俵積田健君(現，松下電工)，在校生の内部英治君，高橋泰岳君，中村理輝君に感謝します．尚，イラストは筆者の次男，浅田龍による．

参考文献

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

[2] R. S. Sutton. “Special issue on reinforcement learning”. In R. S. Sutton(Guest), editor, *Machine Learning*, Vol. 8, pp. –. Kluwer Academic Publishers, 1992.

[3] 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.

[4] 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.

[5] M. Asada, S. Noda, and K. Hosoda. Non-Physical Intervention in Robot Learning Based on LfE Method. In *Proc. of Machine Learning Conferen Workshop on Learning from Examples vs. Programming by Demonstration*, pp. 25–31, 1995.

[6] M. Asada, E. Uchibe, S. Noda, S. Tawaratsumida, and K. Hosoda. “Coordination Of Multiple Behaviors Acquired By Vision-Based Reinforcement Learning”. In *Proc. of IROS94*, pp. 917–924, 1994.

[7] 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.

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

[9] M. Asada, S. Noda, S. Tawaratsumida, and K. Hosoda. “Purposive Behavior Acquisition for a Real Robot by Vision-Based Reinforcement Learning”. *Machine Learning*, Vol. 12, pp. ??–??, 1996.

(あさだ みのる 大阪大学工学部)