学習によるロボットの行動獲得
―サッカーへの適用とロボカップへの取り組み―

鈴木昭二* 浅田稔†

移動ロボット，強化学習，行動獲得，サッカー，ロボカップ

1 はじめに

ロボットが経験を通じて合目的な行動を獲得することは可能だろうか．ある環境内にいるロボットが，試行錯誤を繰り返すうちに，徐々にある目的に沿うように自分自身の行動を選択しタスクを遂行できるようになるだろうか．従来，ロボットの行動決定は，環境の観測，観測情報の解析，行動計画の立案，行動の実行といった段階を踏んで行われ，しかもそれぞれの部分は人間の与えたプログラムにより処理されていた．そのため，環境やタスクがより複雑になった場合にはそれに対処するようプログラムを作り直さなければならなかった．それに対し我々は，ロボット自身が環境との相互作用を通じてタスク遂行のための行動を獲得する手法について研究を行ってきた[2][3]．

実環境内でロボットが行動する際には，ロボットの内部ではおよそ以下の処理が繰り返される．

1. ロボットは自身の持つセンサを通じて環境の様子を観察する．
2. 観察結果に応じてロボットは行動を起こす．
3. ロボットが行動を起こすと，センサの視野が変化するために観察される環境の様子は変化する．
4. 変化した観察結果に応じて，ロボットは再び行動を起こす．

この時，ロボットは観察により環境から情報を得て，行動することにより環境へ働きかけることから，ロボットと環境の間には何らかの相互作用があると考えられる．ロボットはこの相互作用の中で行動選択を行っており，我々はこれらの関係を記述することによりロボットがタスクを遂行するための行動を獲得できると考えている．

本稿では，移動ロボットに強化学習を適用しロボットと環境の相互作用を通じて合目的な行動を獲得させる手法について述べる．また，提案手法を移動ロボットに適用し，サッカーにおけるシュート行動を獲得させることによりその有効性を示す．更に，ロボットによるサッカー大会であるロボカップ[6][5]においても適用してみたのでその結果について述べる．

2 ロボットのシステム構成

我々はサッカーを例にとり強化学習によるロボットの行動獲得の実現を試みた．ここでは，具体的な手法を述べる前に我々のロボットシステムを紹介する．

*すずき・しょうじ 1993年筑波大学大学院博士課程工学研究科終了．博士(工学)．現在大阪大学大学院工学研究科知能・機能創成工学専攻助手．移動ロボットの研究に従事．日本ロボット学会，日本機械学会，IEEE R&AおよびCS会員．

†あさだ・みのる 1982年大阪大学大学院基礎工学研究科後期課程修了．現在，大阪大学大学院工学研究科知能・機能創成工学専攻教授．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）．日本ロボット学会などの会員．

図1にロボットの写真を載せる．本体は市販のラジコン自動車を用い，これにCCDカメラと画像送信機を搭載している．移動機構として左右の車輪が独立したモーターで駆動されるPWS (Power Wheeled Steering) を採用しており，前後進や左右回転の動作ができる．また，ボールを蹴るための道具としてロボットの前面に板が取り付けられている．

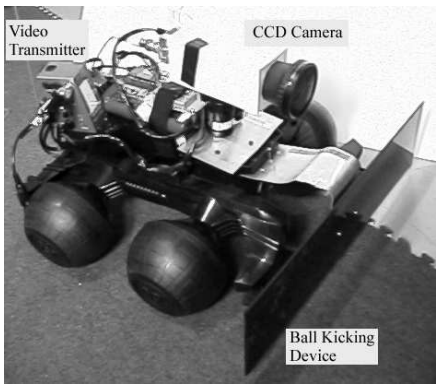


Figure 1: ロボット

図2にロボット制御システムの構成を示す．制御用のホストコンピュータとして PC (Gateway2000 M6-200) を用い，この上で画像処理とロボットへの動作指令の生成を行う．ホストコンピュータとラジコンとのインタフェース部分は，UPP付きV25CPUボードを用いて自作した．

ロボット上のカメラで捉えられた画像は画像送信機によりホストコンピュータに送られる．ホストコンピュータ上では画像処理を行い必要な情報を取り出し，これに対してロボットの動作を決定する．現在は，色の識別による物体の認識を行っている．ラジコンとのインターフェース部はホストコンピュータからの動作指令に従いロボットを無線操縦するための信号を生成する．

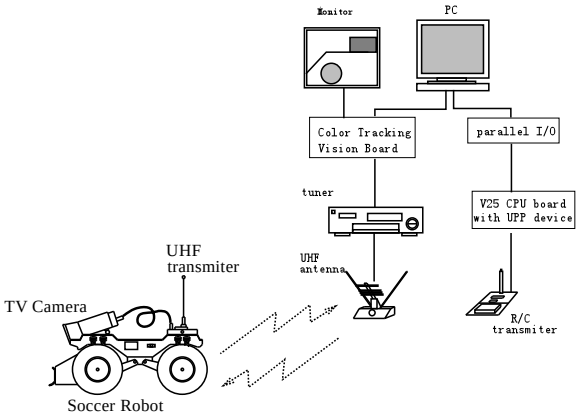


Figure 2: ロボットのコントローラー

3 強化学習の適用による移動ロボットの行動獲得

3.1 強化学習の枠組み

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せずにロボットの合目的な行動の獲得を可能にする点にある．強化学習の基本的な枠組みでは，ロボットは認識した環境の状態に対し行動を起こし，それに対して環境の状態が変化し，更にそれを認識して新たな行動を起こすという処理を繰り返す．この時，ロボットの取った行動に対し，与えられたタスクの達成度合に応じて報酬が与えられる．以上の過程を繰り返すうちにロボットは徐々に高い報酬を得ることのできる行動を選択するようになり，最終的には与えられたタスクを遂行する目的行動を獲得する (図3参照)．

最も良く利用される強化学習法としてQ学習[1]がある．Q学習では，ロボットの行動は行動価値関数により評価され，ロボットは行動価値関数の値を最大にするよう試行錯誤を繰り返す．

今，環境の状態を表す状態集合を$\boldsymbol{S}$，ロボットがとることのできる行動集合を$\boldsymbol{A}$で表し，現在の状態をs，その時にロボットがとった行動をa，次の状態をs'と表す．また，それぞれの状態・

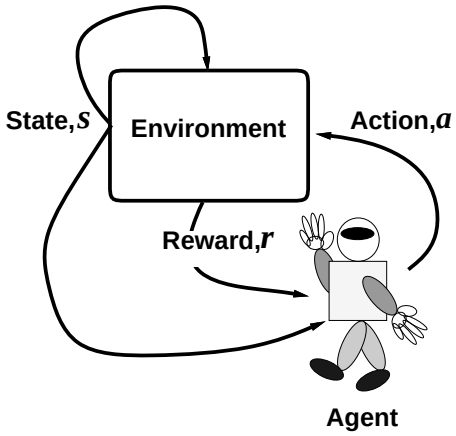


Figure 3: 強化学習によるロボットと環境の相互作用のモデル

行動のペア (s, a) に対し，報酬 $r(s, a)$ を定義する (図3 参照)． この時，状態 s で行動 a をとる行動価値関数 $Q(s, a)$ は，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．

3.2 強化学習による行動の獲得

3.2.1 Q学習によるシュート行動の獲得

ロボットの具体的なタスクとしてサッカーにおけるシュート行動を取り上げた．図1のロボットが図4に示すボールとゴールのみがある簡単な環境下で Q 学習によりシュート行動を獲得する．ボールとゴールは単色に塗られているものとし，ロボットは自身のカメラによりこれらを容易に識別できるものとする．

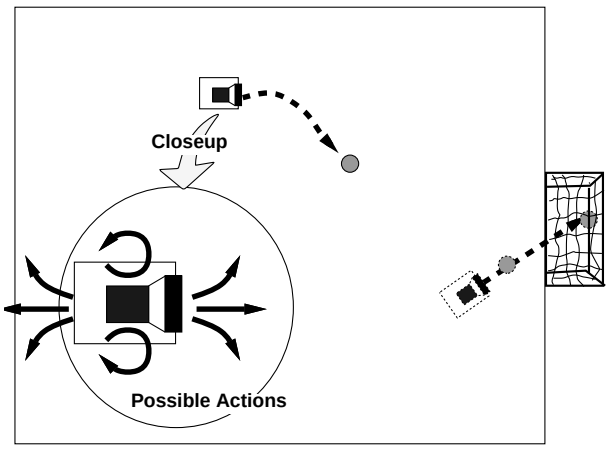


Figure 4: シュート行動と環境

Q 学習を適用するためには，我々が採用した状態の分け方を図5に示す．ボールについては，画像上での大きさ (ボール半径: 大，中，小) と位置 (重心の水平軸上の位置:左，中央，右) 及び，観測されない場合の「右に消えた」,「左に消えた」の 2 状態，ゴールについては，ボールと同様の大きさ (垂直軸方向の長さ)，位置 (水平軸上の座標中心) に加えて向き (ゴールバーの傾き:右向き，正面，左向き) 及び，観測されない場合の「右に消えた」,「左に消えた」の 2 状態を設定し，これらの組合せにより，

- ボール，ゴールともに観測されている場合:3^5(ボール，ゴールの状態数)=243,
- ボールのみがみえている場合:3^2(ボールの位置，大きさ) ×2 (ゴールの消えた方向) =18,
- ゴールのみがみえている場合:3^3(ゴールの位置，大きさ，傾き) ×2 (ボールの消えた方向) =54,
- ボールもゴールもみえていない場合:2(ボールの消えた方向) ×2(ゴールの消えた方向) =4

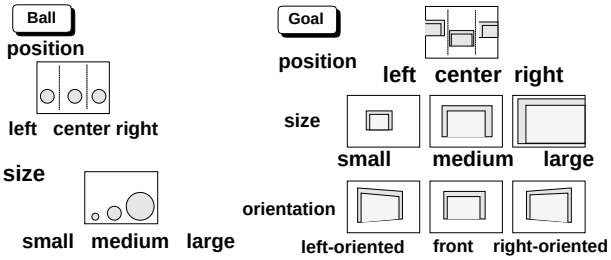


Figure 5: ボールとゴールの状態集合

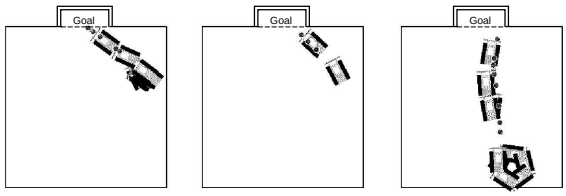
の合計 319 の状態を定義した.

行動集合$\boldsymbol{A}$に関しては，個々のモーターに対する正回転，停止，逆回転コマンドを組合せることにより，9 通りの行動 (直進，左右の回転，右左折前進，後退，右左折後退など) を定義した. なお，ここではモーターの回転速度は一定としており，速度変化はないと仮定した.

3.3 実験と結果

シミュレーションと実ロボットによる実験を行った. 一般に Q 学習は学習が収束するまでに時間を要するため，まず，シミュレーションを行った. 次にシミュレーション結果を実ロボットに移植し検証した.

シミュレーションは SGI のインディゴ Elan (CPU:R4000) を用いて行った. 学習の収束時間は約一日であった. シミュレーション中のロボットの行動例を図 6に示す. 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. 図 6(a) は γ の値が大きい場合で，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. (b) は γ の値が小さい場合で，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する行動となる. (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. シミュレーションにより獲得したシュート行動の成功率は 70%であった.



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

Figure 6: シミュレーション中の行動例

実ロボットによる場合は，シミュレーションに比べ画像にノイズが多い，車輪が滑る，ボールが思わぬ方向によく転ぶことなどが起こる. しかしながら，状態の分け方を粗くしているため，これらの影響は少ない. 但し，粗いゆえにゴール直前での微調整が効かないため，シュート率はシミュレーション時より劣る.

4 強化学習により獲得した行動の実世界への適用－ロボカップへの参加－

4.1 ロボカップ 97 の概要と取り組み

ロボットワールドサッカー 1997(ロボカップ 97) は人工知能国際会議 (IJCAI-98) と併せて 1997 年 8 月 23 日より 28 日にかけて名古屋で行われた. ロボカップはサッカーを行うロボットの実現を通じた実世界で作業を行うエージェントチームを実現するための技術開発を目的とする.

ロボカップ 97 では以下の部門が開催された.

シミュレーション部門 仮想のサッカー競技場 (サッカーサーバー) 上でソフトウェアエージェントにより構成されたチーム同士が対戦を行う. 29 チームが参加した.

実機ロボット部門 実機のロボットを用いた対戦. ロボットの大きさにより小型ロボット部門と中型ロボット部門が開催された. 参加

チームはそれぞれ4チームと5チーム．また，個体のロボットによる技能披露のための特殊技能部門も開催された．

ロボカップについての詳細とロボカップ97の結果に関しては[5] [7][8] を参照していただきたい．

我々は，Q学習により獲得された行動が実験室以外の環境でどの程度有効であるかを確かめるためロボカップ97の中型ロボット部門に参加した．我々は，実世界で作業を行うエージェントチームを実現するためには，個々のロボットが実環境に適応した行動を獲得する能力を持つ必要があると考える．そのための第一段階として，強化学習によるロボットの行動獲得の有効性と限界を確かめる必要がある．

中型ロボット部門のルールはおよそ次のものである．

1. ロボットの大きさと形は直径50cmの円内に収まること．高さの制限はなし．
2. 試合は前半5分，後半5分で行う．ハーフタイムを10分とする．
3. 一チームは最大5台のロボットで構成される．
4. 競技場は大きさは約5mX8mで卓球台9枚分と等しい．競技場の周囲は壁で囲割れている．
5. ボールは赤，2つのゴールはそれぞれ青と黄に塗られる．
6. ロボット同士の衝突回避については対戦チーム同士の合意により試合毎に対処する．

また，試合を円滑に進めるための特別なルールとして，全てのロボットがボールを見失った時など試合が膠着した場合は審判がボールを置き直すことが許された．図7は中型ロボット部門の競技の様子である．

Figure 7: 中型ロボット部門の試合風景

4.2 試合の結果

我々のチームは最終的には5試合を行った．結果は以下の通りである．

試合日	試合の種類	得点	結果
8/25	予選1	1-0	勝
8/26	予選2	2-2	引分
8/27	親善試合1	1-0	勝
8/28	決勝	0-0	引分
8/28	親善試合2	0-1	負

Table 1: 試合の結果

4得点のうち予選2の2得点は相手の自殺点であるため，実質は5試合を通じて2得点しか挙げることができなかった．この2得点はそれぞれ，

- 予選1の得点はキックオフゴールである．
- 親善試合1の得点は，審判が置き直したボールの位置が，たまたま相手ゴール近くにいる味方ロボットの正面であったためである．

ことから，得点に結びついたのは非常に数少ない状況下であったと言える．このことは，3.2で行った状態分けは試合中にロボットが遭遇する状態のごく一部であったことを示している．

図7に示す通り，実際の競技場はかなり簡単化された環境であり，しかもボールやゴールは色分けされており物体の認識がしやすいように配慮されているため，我々の実験環境と共通する点が多いと考えられる．．そのため，我々は実験を通じて得られた結果がそのまま適用できることを期待したが，実際には様々なノイズのためになかなかロボットが思うように動いてはくれなかった．

1. ラジコンのノイズ
   ラジコン電波に乗ったノイズのためにモータが振動した．その状態のまま試合を続けたため過大な電流がながれつづけモータが発熱し発煙した．
2. 画像転送のノイズ
   ロボットからホストコンピュータへはカラー画像を伝送するが，会場の電波状態が良くなかったため白黒映像になってしまうことがあった．ロボットは色によりゴールやボールを認識しているので，白黒画像ではこれらを発見することができない．
3. 照明の影響
   競技場にある程度の広さがあるため，競技場内の照明は完全には一様でない．そのために場所により色が異なって見える場合があり，ゴールやボールを見失うことが多かった．

実験環境との大きな違いは，他のロボットの存在である．競技中は味方と相手を含めて10台前後のロボットが競技場内に存在する．今回は，個々のロボットのシュート行動のみを実現したため，味方ロボットの認識と連係プレーに関しては手つかずであった．また，相手チームのロボットに関しては，これを障害物として認識し障害物回避行動を学習させることにより対処した．相手ロボットとの衝突を避ける行動とシュート行動の統合は[4]の手法を適用した．

5 まとめ

本稿では，ロボットと環境の相互作用に着目し，移動ロボットが強化学習により合目的な行動の獲得を行う手法を示した．タスクとしてサッカーにおけるシュート行動という簡単なものを与え，シミュレーションと実ロボットにより行動獲得できることを確認した．しかしながら，実験室においては一応の成果をおさめたものの別の環境下では獲得した行動は有効に働かなかった．原因としては，環境観察の不確かさの問題もあるが，ロボットの学習時に与えた状態集合にも問題があると考えられる．現在は，これをロボット自身が生成するための手法について研究を行っている[9][10][11]．

従来のロボットの研究においては，ロボットに何を与えることによりどう行動するのかを明らかにすることが中心テーマであった．これからは，ロボットが行動を獲得するために必要な基本的な枠組を明らかにしていくことが重要なテーマである．

謝辞

大阪大学工学研究科知能・機能創成工学専攻浅田研究室の学生諸君，特にロボカップ97に積極的に取り組んでくれた，内部英治，高橋泰岳，中村理輝，三島千寿子，石塚宏，加藤龍憲の諸君に深く感謝する．

References

[1] C. J. C. H. Watkins: “ Learning from delayed

rewards", PhD thesis, King's College, University of Cambridge, May 1989.

[2] 浅田: "「特集ロボカップ」第3章ロボットプレーヤの感覚と学習", bit, Vol.28, No.5, pp.37-43,1996.

[3] 浅田稔・野田彰一・俵積田健・細田耕: "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, No.1, pp.68-74,1995.

[4] 内部英治・浅田稔・野田彰一・細田耕: "視覚を有する移動ロボットの強化学習による複数タスクの達成", 機械学会ロボティクス・メカトロニクス講演会 95 予稿集,1995.

[5] http://www.robocup.org/RoboCup/RoboCup.html

[6] H.Kitano, M.Asada, Y.Kuniyoshi, I.Noda, and E.Osawa: "RoboCup: The Robot World Cup Initiative" Proc. of IJCAI-95 Workshop on Entertainment and AI/ALife, 1995.

[7] http://www.robocup.org/RoboCup/RoboCup97.html

[8] 松原仁・北野宏明・浅田稔・野田五木樹・鈴木昭二: "RoboCup-97 報告", 情報処理学会誌, vol.38, no.12, 1997

[9] E.Uchibe, M.Asada, and K.Hosoda: "Vision Based State Space Construction for Learning Mobile Robots in Multi Agent Environments", Proc. of Sixth European Workshop on Learning Robots(EWLR-6), pp.33-41, 1997.

[10] E.Uchibe, M.Asada, and K.Hosoda: "Strategy Classification in Multi-agent Environment – Applying Reinforcement Learning to Soccer Agents", ICMASS'96 Workshop on RoboCup Workshop: Soccer as a Problem for Multi-Agent Systems, 1996.

[11] Y. Takahashi, M.Asada, S.Noda, and K.Hosoda: "Sensor Space Segmentation for Mobile Robot Learning", ICMASS'96 Workshop on Learning, Interaction and Organizations in Multiagent Environment, 1996.