日本ロボット学会誌　Vol. 13　No. 1, pp.68〜74, 1995

学術論文

視覚に基づく強化学習によるロボットの行動獲得

浅田 稔*　野田 彰一 *　俵積田 健 *　細田 耕 *

Purposive Behavior Acquisition for a Robot by Vision-Based Reinforcement Learning

Minoru Asada*, Shoichi Noda* , Sukoya Tawaratsumida *　*and* Koh Hosoda*

We propose a method which acquires a purposive behavior for a mobile robot to shoot a ball into the goal by using the Q-learning, one of the reinforcement learning methods. A mobile robot (an agent) does not need to know any parameters of the 3-D environment or its kinematics/dynamics. Information about the changes of the environment is only the image captured from a single TV camera mounted on the robot. Image positions of the ball and the goal are used as a state variable which shows the effect of an action taken by the robot during the learning process. After the learning process, the robot tries to carry a ball near the goal and to shoot it. Computer simulation is used not only to check the validity of the method but also to save the learning time on the real robot. The real robot succeeded in shooting a ball into the goal using the learned policy transferred to it.

Key Words: Vision-Based Reinforcement Learning, Q-Learning, Behavior Acquisition, Soccer Robot, State-Action Deviation

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[?]．この問題に対処するために，Brooks[1]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[2][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(**Fig.1**参照)．

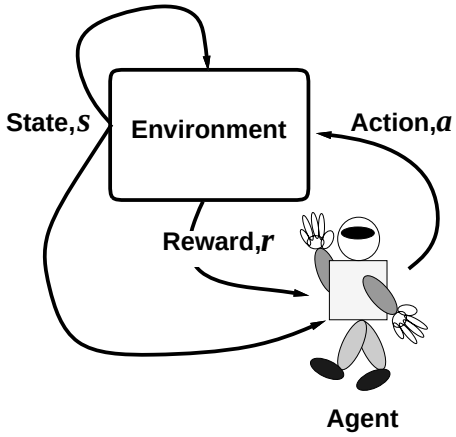


Fig. 1　The basic model of robot-environment interaction

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．ところが，従来の強化学習の研究では，学習の収束を早めるための理論的考察や，複数タスクへの拡張法[5]を論議しているものが多い．しかも，そのほとんど

原稿受付　1994年3月14日
*大阪大学工学部
*Faculty of Engineering, Osaka University

がコンピュータシミュレーションによる結果しか提示しておらず[6]〜[8]，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．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[9]．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．

我々が知る限りでは，実ロボットに適用した例として，以下の二つが挙げられる．一つは，Maes and Brooks[10]の6本脚ロボットの歩行実験で，強化学習に類似した手法を用いているが，状態数が少なく，反射的で簡単な実験例である．もう一つは，Mahadevan and Connel[11]による，箱押し作業ロボットである．実環境での学習に多大な時間を要するので，箱押し作業を事前に「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なければならない．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．

このように，コンピュータシミュレーションだけの実験では，状態空間の構成法が，過剰に理想的であったり，実際のロボットの実験では，多大な学習時間を要するため，簡単なサブタスクに分解し，反射的な行動を学習させる場合が多い．実環境で，自律的なエージェントを実現するための強化学習の役割を明確にするためには，より動的で複雑な環境で目的行動を獲得する応用例が必要である．

本論文では，サッカーロボットを例として，強化学習を実際のロボットに適用する際の問題点を解決し，ロボットがシューティング行動を実現する手法について述べる．本研究の最大の特徴は，世界に関する知識，例えば，ボールやゴールの3次元位置や大きさ，ロボットの運動学・動力学的パラメータなどの知識を一切必要とせずに，ボールをゴールにシュートする行動を獲得することである．ロボットが利用できる情報は，TVカメラから得られるボールやゴールの映像情報だけであり，それらはロボットが選択した行動により変化する．ロボットの行動は前後進や回転であるが，それらの物理的意味もロボットは知る必要がない．我々の知り得る限り，本研究で示すような実時間視覚を用いた強化学習によるロボットの目的行動獲得の研究は，ほとんど見当たらない．

視覚に基づく強化学習では，視覚情報を基本として状態空間を構成するので，画像上で識別可能な空間と，ロボットの3次元空間とは必ずしも一致しない．TVカメラの近傍では分解能が高く，逆に遠方では低い．これに対し，ロボットの行動は元の3次元世界で，ほぼ同じ量の運動として実現されるので，ロボットによる行動と状態遷移が1:1に対応するとは限らず,「状態と行動のずれ」が生じる．ここでは，ロボットの実際の行動を行動要素として定義し，状態変化を伴うまでの同一行動要素の集合を行動として再定義することにより，このずれ問題に対処した．

学習を効率的に進めるためには，多大な時間を要する学習を実ロボット上で実施するのは，好ましくない．そこで，よく利用されるコンピュータシミュレーションを利用(例えば，文献[12])して，強化学習を実施し，その結果得られる「各状態に対する最適行動の政策」を実際のロボットに移植する．これにより，シミュレーションと実世界とのギャップが明らかになり，システムの改良に役立つ．

以下，次章では，強化学習の基本的枠組み及び，実験で用いたQ-学習について簡単に説明する．次に，タスクや基本的な仮定を示し，学習の方法を説明する．コンピュータシミュレーションと実機による実験結果を提示し，最後に，考察及び今後の課題を述べる．

2. 強化学習の基本的枠組みとQ-学習

2.1 強化学習の基礎

ロボットは環境の状態を表す状態集合$\boldsymbol{S}$を識別可能で，環境に対し，行動集合$\boldsymbol{A}$の中の一つの行動をとることができる．このとき，環境はマルコフ過程としてモデル化され，現在の状態とロボットがとった行動により確率的な状態遷移を行なう．現在の状態をs，その時，ロボットがとった行動をa，次の状態をs'とすると，$T(s,a,s')$は，そのときの状態遷移確率を表す．また，それぞれの状態・行動のペア(s,a)に対し，報酬$r(s,a)$が定義される(**Fig.1**参照)．

一般的な強化学習の問題は，無限時間に対する報酬の減衰総和を最大化する政策を発見することである．政策fは，状態集合$\boldsymbol{S}$から行動集合$\boldsymbol{A}$へのマッピングである．この総和は「積算報酬(return)」と呼ばれ，

$$\sum_{n=0}^{\infty}\gamma^n r_{t+n}, \qquad (1)$$

と定義される．ここで，r_tは，各状態で政策fをとった時の時刻tにおける報酬を表す．γは減衰係数を表し，将来の報酬がどの程度行動価値に影響を与えるかを制御し，通常，1より小さい値を持つ．

遷移確率と報酬の分布が与えられれば，動的計画法の枠組みで最適政策を求めることができる[13]．

2.2 Q-学習

実際のロボットの環境では，遷移確率と報酬の分布が完全に既知である場合は少なく，試行錯誤的に探索しながら，最適な行動を学習することが考えられる．以下では，そのような探索的な学習法である「Q-学習」[14]について説明する．

状態sで，行動aをとり，それ以降最適政策をとった時の積算報酬の期待値もしくは「最適行動価値関数」を$Q^*(s,a)$とする．これは，再帰的に

$$Q^*(s,a)=r(s,a)+\gamma\sum_{s'\in S}T(s,a,s')\max_{a'\in\boldsymbol{A}}Q^*(s',a') \qquad (2)$$

と定義される．最初は，遷移確率Tや報酬rが未知なので，オンラインで逐次的に「行動価値」Qを更新する．初期値として任意の値(通常は0)から始めて，行動がとられる毎に，Qの

値を

$$Q(s,a) \Leftarrow (1-\alpha)Q(s,a) + \alpha(r(s,a) + \gamma \max_{a' \in \boldsymbol{A}} Q(s',a')) \tag{3}$$

として更新する．ここで，rは，状態sで行動aをとった時の報酬，s'は次状態を表す．またαは，学習率で0と1の間の値をとる．本研究では，以下の1ステップQ学習アルゴリズムを用いた．

(1) Q ← 行動価値関数の初期値を代入(通常はゼロ)．

(2) 現在の状態sに対し，政策fにしたがった行動aを選択し(任意でもよい)，実行する．

(3) $Q(s,a)$の更新:

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

ここで，s', rは，それぞれ，次状態，即座の報酬を表す．

(4) 政策fの更新:

$$Q(s,a) = \max_{b \in \boldsymbol{A}} Q(s,b) \quad \text{となるような} \quad a \rightarrow f(s)$$

(5) 行動価値関数及び政策が収束状態に達したと判断されない時，ステップ(2)に戻る．

3. タスクと仮定

タスクとして，**Fig.2(a)**に示すように移動ロボットがボールをゴールにシュートすることを考える．問題は，このような行動を獲得する手法を開発することである．問題を簡単化するために，フィールド内にはゴールやボール以外のものは存在しないとする．**Fig.2(b)**に，実際に用いた移動ロボット，ボール及びゴールを示す．

既に述べたように，ロボットは，ゴールの大きさやその3次元位置，ボールの大きさ，重さ及びその3次元位置，焦点距離や傾き角などの全ての内部，外部カメラパラメータ，そしてロボット自身の運動学，動力学パラメータ及びその物理的意味について全く知る必要がない．与えられる情報は，TVカメラから得られるボールやゴールの映像情報だけである．

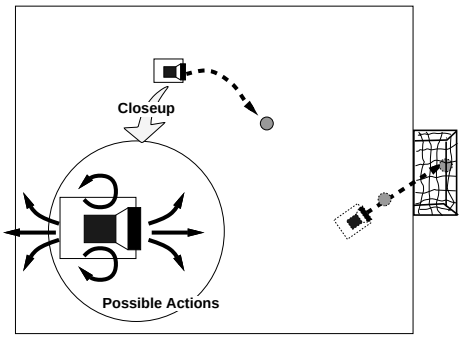
4. 状態空間及び行動空間の構成

Q学習を適用するために，状態空間，行動空間などを構成しなければならないが，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である．しかしながら，実際のセンサーから得られる状態空間と実際のロボットの行動空間とは，明確に対応するとは限らず,「状態と行動のずれ」問題が生じる．以下では，実世界に対応する状態空間と行動空間の構成法を述べ，その後,「状態と行動のずれ」問題の解決法を示す．

4.1 各空間の構成

(a) 状態集合Sの構成

ロボットが環境の状態を感知できる唯一の情報はTVカメラから得られるボールやゴールの像だけである．精度良く最適な行動を選択するためには，状態空間の分解能は高い方がよい．



(a) The task of shooting a ball into the goal

(b) A picture of the radio-controlled vehicle with a ball and a goal

Fig. 2 A task and our real robot

しかし，画像処理のノイズや，学習時間が状態空間の大きさの指数関数オーダーである[?]ことを考えると，粗い方がよい．そこで，**Fig.3**に示すように，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)$\times$2(ゴールの消えた方向) =18,
- ゴールのみがみえている場合:3^3(ゴールの位置，大きさ，傾き)$\times$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) $\times$2(ゴールの消えた方向) =4

の合計319の状態からなる状態空間を構成した．

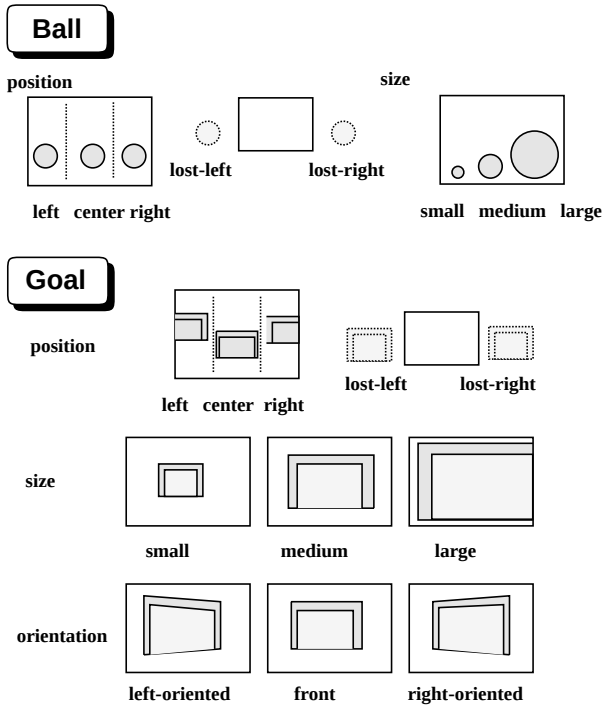


Fig. 3 The ball substates and the goal substates

(b) 行動集合Aの構成

ロボットは，二つの独立のモーターで駆動されるPWS (Power Wheeled Steering)を採用しており，それぞれに個別のコマンド指令を送ることができる．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) が選択できる．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．

(c) 報酬及び減衰係数γ

報酬を各状態に対して，割り当てることも考えられるが，厳密に報酬を割り当てない限り，行動価値関数の更新過程で，多くの極大値が発生し，学習が収束しない[15]．そこで，ボールがゴールに入った時$r = 1$，それ以外は$r = 0$を報酬として与える．また，減衰係数に関しては，$\gamma = 0.8$とした．

4.2 「状態・行動空間のずれ」問題

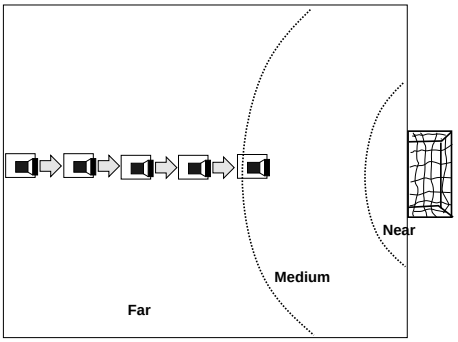


Fig. 4 A state-action deviation problem

画像情報の特性として，近傍の分解能に比べて，遠方の分解能が粗いことが挙げられる．これに対し，行動は，3次元環境に対して，常に同程度の運動を示すので，画像情報から識別される状態と3次元環境での行動がずれる可能性がある．**Fig.4**は，この状態を表しており,「ゴールが小さく見える(遠方)」状態に対応するフィールドの領域は,「ゴールが大きく見える(近傍)」状態に対する領域よりかなり広い．このことは状態遷移のばらつきが大きいことを意味し，正しく学習が収束することを阻む．**Fig.4**の場合，ロボットが直進行動をとった時，頻繁に同じ状態に戻ることになり，最適な政策を得ることは困難である．

そこで，行動空間を以下のように再構成した．上で定義した行動$a \in \boldsymbol{A}$を行動要素とし，ロボットは，状態が変化するまで，同じ行動要素を実行する．状態変化が起きたとき，同一行動要素の系列を一つの行動とみなして，行動価値関数の更新式(3)を適用する．**Fig.4**の場合，ロボットは直進行動要素を何度か繰り返し,「ゴールが小さく見える」状態から「ゴールが中位の大きさに見える」状態に遷移する．このとき初めて，行動価値関数を更新する．

5. 実験

実験は二つの部分からなる．最初に，コンピュータシミュレーションで最適政策fを学習を通して獲得する．次に，最適政策fを実際のロボットに移植し，これに従いロボットを制御する．コンピュータシミュレーションの効用は，単に，アルゴリズムの検証だけでなく，実機による学習コストの軽減も含む．コンピュータシミュレーションは，完全には実世界を模擬できない[16]ので，政策の移植は，システムの改良，実機とシミュレーションの違いの発見などに役立つ．

5.1 シミュレーション結果

コンピュータシミュレーションは，**Fig.2(b)**に示した実際の移動ロボット，ボール，ゴールの大きさや，重さ，床面の摩擦，ロボットとボールの反発係数，カメラパラメータ(焦点距離や，傾き角など)などを実測し，それらを反映するようにした(実数値は次節参照)．但し，厳密に計測しているわけではないので，あまり正確とはいえない．

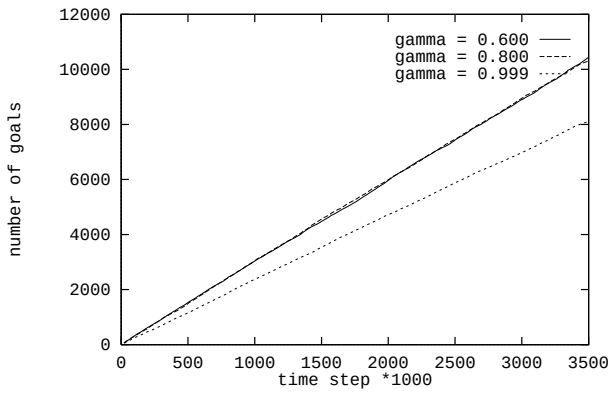
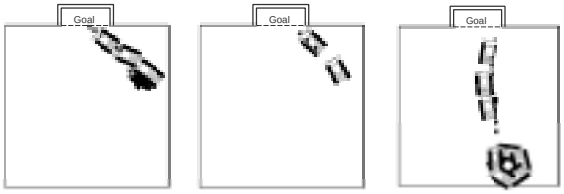


Fig. 5 Number of goals in terms of γ

学習の収束時間は，SGIのインディゴElan(CPU:R4000)で，約一日である．**Fig.5**は，減衰係数γによる，学習中のシュート数の違いを示している．縦軸が累積のシューと数で，横軸は，行動のサイクルタイム(実ロボットでは，約33[ms])によるステップ数を表している†．学習後は，γの値が大きい(0.999)時は，小さい値(0.8もしくは0.6)の時よりシュート回数が少ない．これは，γの値が大きい時，どんなに時間を掛けてもゴール時の報酬が，あまり変わらないのに対し，小さい値の時は，早くゴールすればするほど報酬が高いからである．但し，小さ過ぎると，ゴールに辿りつく経路を見失うおそれがある．**Fig.6(a,b)**にその例を示す．上部中央にゴールがあり，黒点がボールを示している．(a)では，γの値が大きいので，少しでも確実にシュートするために，よりよい位置に移動してからシュートしているのに対し，(b)では，γの値が小さいので即座にシュートしている．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．



(a) shooting ($\gamma = 0.999$) (b) shooting ($\gamma = 0.6$) (c) finding, dribbling, and shooting

Fig. 6 Some kinds of behaviors during learning process

5.2 実ロボットによる実験

Fig.7に，実際構築したシステムを示す．本システムでは，Inaba[17]による遠隔脳アプローチ(Remote-Brained Approach)を採用した．これは，高度で重装備を必要とする情報処理部(脳)から，実世界で行動する本体(ボディー)を分離することにより，種々の知能ロボットを容易に実現可能にする研究環境を与える．具体的には，ロボット本体上に情報処理装置を置かず，遠隔のホストコンピュータに無線でデータを送信後処理し，結果得られる制御指令も本体に無線で送信するものである．

ロボット上のTVカメラ(Sony社製ハンディカムTR-3，画角は36度に固定)で撮られた画像は，ビデオ送信器でホストのUHF受信器に送られ，パイプライン型高速画像処理装置(MaxVideo200)で処理される．処理の簡単化と高速化のために，ボール，ゴールはそれぞれ赤と青に塗装されている．画像処理，状態識別，行動選択は，ホストCPU(MC68040)上のOS(VxWorks)によって制御される．ホストCPUはイーサネットを介してSunワークステーションに接続されており，プログラム開発などが容易である．既に，**Fig.2(b)**に実際のロボットなどを示したが，それらの主要諸元は以下の通りである．ロボットは，幅0.32[m]，長さ0.52[m]，カメラの高さ0.18[m]，重さ4.19[kg]，最大直進速度1.1[m/s]，最大角速度4.77[rad/s]，ボールは直径0.09[m]，重さ0.07[kg]，ゴールは幅0.45[m]，高さ0.27[m]，ポール幅0.055[m]である．

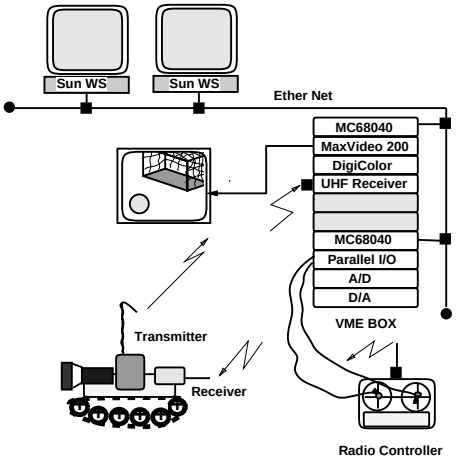
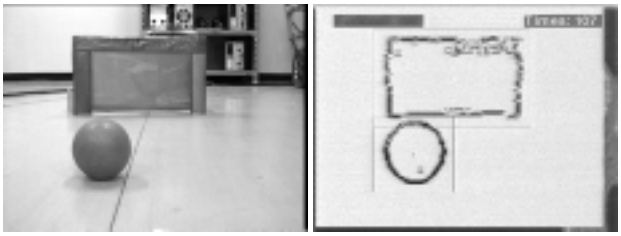


Fig. 7 A configuration of the real system



(a) input image (b) detected image

Fig. 8 Detection of the ball and the goal

Fig.8(a,b)に，ロボットから送信された原画像(実際はカラー画像)とボール及びゴールを検出した画像を示す．処理内容は，色によるボール，ゴール領域の抽出，画像縮小(面積1/16)，エッジ抽出，特徴点配列作成で，検出時間はパイプライン処理により33[ms]要する(詳細は，文献[18]参照)．また，状態識別，行動選択はホストCPU上で行なわれ，約8[ms]要する．よって，サンプリング時間が約33[ms]，遅れが約41[ms]である．行動選択結果は無線コントローラからロボット本体に送信される．

Table 1は，**Fig.9**に示したロボットのシューテイング行動に対する状態識別結果を示す．**Fig.9**では，6つの時刻の環境の

†1回のシュートに対するステップ数が平均350から400と多いのは，学習中でのランダムな行動選択に起因する．但し，行動選択確率$P(s,a)$は以下の分布を用いており，同じパラメータT(ボルツマンマシンの温度パラメータに対応)を用いているので，γによる差が出ている．

$$P(s,a) = \frac{e^{Q(s,a)/T}}{\sum_{a' \in A} e^{Q(s,a')/T}}$$

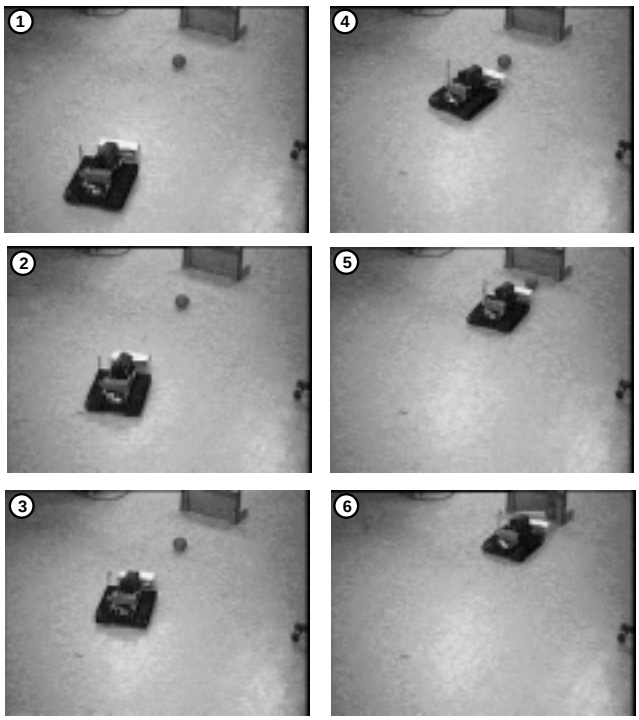


Fig. 9 The robot succeeded in shooting a ball into the goal.

様子を示している．**Table 1**では，33[ms]毎のtime step，ロボットが識別したstate step，ボールに対する状態(左(L)，中央(C)，右(R)，消失(D)と近い(N)，中くらい(M)，遠い(F)の組合せ)，ゴールに対する状態(ボールの状態に加えて，正面(Fo)，左向き(Lo)，右向き(Ro)が加わる)，左右輪に対する制御コマンド(前進(F)，停止(S)，後退(B))，そして識別に失敗した状態数が示されている．識別に失敗した状態には(*)印が付けられている．

Table 1は平均的な状態識別結果であり，驚くべきことに全状態数の7割り弱(67%)しか，正しく識別されていない．間違った識別は，ボールの大きさやゴールの向きの判断を誤ったときが多く，これは，画像処理のエッジ抽出の失敗，ロボットの微小な上下動などが原因と考えられる．ボールとゴールが画像のほぼ中央に位置する時は，ボールの大きさを誤って判断しても，あまり差し支えない．なぜなら，その場合，最適行動は直進だからである．しかしながら，中央にない場合は，距離推定を誤り，失敗する可能性がある．**Table 1**には，表れていないが，ビデオ送信器のノイズで全く状態識別ができないケースが発生している．そのような系統的なノイズが発生している場合は，状況識別を続けて誤り，正しい行動をとれない．しかし，ビデオノイズなどのように発生が独立でランダムである場合は，続けて状況識別を誤る確率が低く，次のステップではほぼ正しい状態識別がなされるのであまり問題がないと考えられる．

6. 検討及び考察

視覚情報は学習中の状態識別や行動評価に欠かせない．特に，大局的な情報を必要とする目標指向型のタスクではより重要である．これまでのコンピュータビジョンのアプローチでは，2次元画像からの3次元再構成問題を中心課題として来ており，様々なタスクに応用できることを期待して，再構成される情報の厳密さを追い求めてきたが，それらは，必ずしも必要でない．個々のタスクに応じて，それらの記述を変換する必要がある場合，そのコストは，実世界で行動するロボットにとって，不利となる場合がある．逆に，タスクやロボットの行動に根ざした環境の記述を最初から作成すべきと考えられる．この意味で，強化学習で得られる，行動価値関数は，状態と行動の最適なマッピング情報が盛り込まれ，タスクもしくはロボットの行動に基づく「環境地図」と考えられる．

ロボットにとって，行動価値に基づく「環境地図」が類似しておれば，それは，環境の類似性を意味する．例えば，障害物発見や回避をタスクとした場合，机や椅子が乱雑に配置されたオフィースシーンも，岩などを含む屋外シーンも識別する必要がない．これまでの幾何情報の再構成を主眼としてきたアプローチでは，これらを同一視することは困難と考えられる．

本論文で提案した「視覚に基づく強化学習」では，画像情報から得られる状態の分解能は，非常に粗い．これは，状態数を軽減させて学習時間を速める効果と，ノイズによる画像処理の誤差を吸収する効果がある．実験で示したように，ノイズの影響などで，粗いマッピングすら間違う場合がある．それが系統的なノイズ発生により連続して続くという状況を除けば，次の

Table 1 State-Action data

time step	state step	state: ball	state: goal	action: L	action: R	error
1	1	(C,F)	(C,F,Fo)	F	F	
2	2	(R*,F)	(C,F,Fo)	F	F	1
3	3	(D*,D*)	(C,F,Ro*)	B	B	3
4	4	(C,F)	(C,F,Lo*)	B	S	1
5	5	(C,F)	(C,F,Fo)	F	F	
6		(C,F)	(C,F,Fo)	F	F	
7		(C,F)	(C,F,Fo)	F	F	
8		(C,F)	(C,F,Fo)	F	F	
9	6	(C,F)	(C,F,Ro*)	B	S	1
10	7	(C,F)	(C,F,Fo)	F	F	
11	8	(C,F)	(R,M,Fo)	F	F	
12	9	(R,F)	(R,M,Fo)	F	F	
13	10	(R,M*)	(R,F*,Lo*)	F	B	3
14	11	(L*,F)	(R,M,Ro*)	F	S	2
15	12	(L*,F)	(R,M,Fo)	F	S	1
16	13	(R,M)	(R,M,Fo)	S	B	
17	14	(C,M)	(C,M,Fo)	F	F	
18	15	(L,M)	(L,M,Fo)	S	F	
19	16	(L,N)	(L,M,Fo)	B	S	
20		(L,N)	(L,M,Fo)	B	S	
21	17	(L,M*)	(L,M,Fo)	S	F	1
22	18	(L,N)	(L,M,Fo)	B	S	
23		(L,N)	(L,M,Fo)	B	S	
24	19	(C,N)	(C,M,Fo)	F	B	
25	20	(C,M)	(C,M,Fo)	F	F	
26		(C,M)	(C,M,Fo)	F	F	
27	21	(C,M)	(C,N,Fo)	F	S	
28	22	(C,M)	(C,M*,Lo*)	F	S	2
29	23	(C,M)	(C,M*,Ro*)	S	B	2
30	24	(C,F)	(D,D,D)	F	S	

ステップで，連続して状態識別を誤る確率が低く，システムのロバスト性を示していると考えられる．

7. おわりに

目的行動を獲得する自律ロボットの実現するための第一段階として,「視覚に基づく強化学習によるロボット行動獲得」法を提案し，コンピュータシミュレーションによる学習法と，それを実機に移して，タスクを遂行する実験でその有効性を示した．動的な環境でのロボットのタスクに，強化学習を適用する問題点として「状態と行動のずれ問題」を指摘し，視覚センサーから構成される状態空間に対応する行動空間を，実際のロボットの行動要素の系列として定義することにより，この問題を解決した．

ここでは，政策の移植を，学習を効率的に実施する手法として用いたが，学習法自体についても，高速化の工夫が必要である．幾つかの高速化手法が提案されている [?] が，それらとの比較も含めた考察については，別稿に譲る．

実機による実験では，ボールがシミュレーションと異なる動きを示す場合があり，シュートが失敗することがあった．現在，実機での学習は行なっていないが，シミュレーションで得られた学習結果を初期値として，学習を行なえば，実機のロボットの学習の効率化だけでなく，類似のタスクに適用する問題 (Scaling Problem) [15] にも関連すると考えられ，現在，実験を計画中である．今後の課題として，複数競技者による協調，競合問題を扱う予定である．

参 考 文 献

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

[2] M. J. Mataric. "Integration of representation into goal-driven behavior based robots". *IEEE J. Robotics and Automation*, Vol. RA-8, pp. –, 1992.

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

[6] S. D. Whitehead and D. H. Ballard. "Active perception and reinforcement learning". In *Proc. of Workshop on Machine Learning-1990*, pp. 179–188, 1990.

[7] 田野, 三上, 嘉数. 「動的環境における多足歩行機械の適応的歩容計画」. 第11回 日本ロボット学会学術講演会 予稿集, pp. 1103–1106, 1993.

[8] 徳本, 三上, 嘉数. 「強化学習を用いた周期的歩行運動の獲得」. 第11回 日本ロボット学会学術講演会 予稿集, pp. 1107–1110, 1993.

[9] 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.

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

[11] 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.

[12] 阪口, 藤田, 升谷, 宮崎. "「動体を扱うロボットの運動計画と制御」. 日本機械学会論文集C編, Vol. 59, No. 567, pp. 3405–3410, 1993.

[13] R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.

[14] C. J. C. H. Watkins. *Learning from delayed rewards"*. PhD thesis, King's College, University of Cambridge, May 1989.

[15] J. H. Connel and S. Mahadevan. "Introduction to robot learning". In J. H. Connel and S. Mahadevan, editors, *Robot Learning*, chapter 1. Kluwer Academic Publishers, 1993.

[16] R. A. Brooks and M. J. Mataric. "Real robot, real learning problems". In J. H. Connel and S. Mahadevan, editors, *Robot Learning*, chapter 8. Kluwer Academic Publishers, 1993.

[17] M. Inaba. "Remote-brained robotics: Interfacing ai with real world behaviors". In *Preprints of ISRR'93*, Pitsuburg, 1993.

[18] 俵積田, 野田, 浅田, 細田. 「視覚に基づく強化学習によるサッカーロボットのシューティング行動の実現」". 第4回 ロボットシンポジウム講演会予稿集, 1994.

浅田 稔

1953年10月1日生．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．博士（工学）．電子情報通信学会，情報処理学会，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　（日本ロボット学会正会員）

野田 彰一

1971年1月19日生．1993年大阪大学工学部機械工学科卒業．現在同大学院工学研究科博士前期課程在学中 (電子制御機械工学専攻)．知能ロボットの学習に関する研究に従事．

俵積田 健

1971年4月19日生．1994年大阪大学工学部電子制御機械工学科卒業．現在，同大学大学院工学研究科前期課程在学中(機械工学専攻)．ガラス繊維強化プラスチックの成形に関する研究に従事．

細田 耕

1965年11月9日生．1988年京都大学工学部精密工学科卒業．1993年同大学工学研究科機械工学先攻博士後期課程修了．同年大阪大学工学部電子制御機械工学科助手となり，現在に至る．ロボット工学の研究に従事．博士（工学）．計測自動制御学会の会員．　（日本ロボット学会正会員）