日本ロボット学会誌　Vol. 17　No. 1, pp.118 ～ 124, 1999

学術論文

実ロボットによる行動学習のための状態空間の漸次的構成

高橋 泰岳[*1]　浅田 稔[*2]

Incremental State Space Segmentation for Behavior Learning by Real Robot

Yasutake Takahashi[*1] *and* Minoru Asada[*2]

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. However, there are two major problems in applying it to real robot tasks: how to construct the state space, and how to accelerate the learning. This paper presents a method by which a robot learns a purposive behavior within less learning time by incrementally segmenting the sensor space based on the experiences of the robot. The incremental segmentation is performed by constructing local models in the state space, which is based on the function approximation in terms of the sensor outputs and the reinforcement signal to reduce the learning time. The method is applied to a soccer robot which tries to shoot a ball into a goal. The experiments with computer simulations and a real robot are shown. As a result, our real robot has learned a shooting behavior within less than one hour training by incrementally segmenting the state space.

Key Words: Segmentation, Inter-dependence between state and action spaces, State space construction, Reinforcement learning

1. はじめに

近年，環境とエージェントの相互作用を通して学習する手法として強化学習 [1] や記憶に基づく学習法などのロボット学習 [2] が注目されている．強化学習の枠組みでは，ロボットと環境はそれぞれ離散化され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，それに対し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．

実環境におけるロボットタスクに強化学習を適用する場合，最も困難な問題の一つは状態空間の構成である．従来の研究では殆どがプログラマが状態空間を事前に設計するが，センサ情報を人間が適当に離散化した状態空間がロボットにとって最適なものとなっている保証はない．状態の離散化が粗すぎると，一つの状態に対する最適な行動が獲得されない (「知覚的見せかけ問題」 [3] と呼ばれている)．また不必要に細かく状態を分割すると，本来同一と見なして良い状態の学習が進まないので経験の汎化が期待できず，また学習時間が状態数に応じて指数関数的に増加し，結果として膨大な学習時間を要する．

この問題に対する手法は大まかには関数近似の手法を用いて行動価値関数を学習させる方法と，状態空間を分割して行動価値関数を学習する方法とに分類できる．

前者の研究例として， Boyan et al [4] は格子状世界でのナビゲーションや自動車の山登りというタスクに対する行動価値関数を回帰モデルやニューラルネットを用いて近似させることを試みた．しかし回帰モデルやニューラルネットと動的計画法の組み合わせは最適解への収束性が保証されておらず，シミュレーションによる実験を通してこの手法がロバストではないことを報告している．一方 Sutton [5] や Saito et al [6] は関数近似の手法として CMAC [7] [8] を用いた． Sutton は Boyan et al と同じタスクに適用し，ロバストに学習できることを示した． Saito et al はブラキエーションのタスクに適用した．しかし CMAC はルックアップテーブルを使った関数近似であり，適用する場合量子化の問題が生じる．十分に離散化すると探索空間が膨大になるので，Saito et al は目的の行動知識をもった初期コントローラーを使い目的の動作領域に近い学習空間のみを探索させている．この様に関数近似の手法では基本的に行動価値関数の収束が保証されておらず，何らかの事前知識などによりかなり正しい初期値を設定する必要がある．

センサ空間を分割する手法の研究として， Kröse et al [9],

原稿受付　1998 年 1 月 28 日

[*1]大阪大学工学部
[*2]大阪大学工学部
[*1]Osaka University
[*2]Osaka University

Dubrawski et al [10] はソナーを持った移動ロボットによる障害物回避を行った．移動ロボットが障害物にぶつかったときに負の報酬を与え，ロボットはそれをもとにセンサ空間を分割して，状態空間を構成し，それを用いて障害物の回避行動を獲得した．Ishiguro et al [11] は全方位視覚を持った移動ロボットによるナビゲーションを行った．移動ロボットが壁にぶつかったとき負の報酬を与え，目的地に到着したとき正の報酬を与えた．ロボットは与えられた報酬の分布によりセンサ空間を分割し，状態空間を構成する．しかしこれらの手法は基本的に報酬信号のみを状態分割の基準にしているため，タスクを達成したときのみ報酬が与えられるので，報酬から遠いセンサ空間の分割が遅れ，状態空間の構成のための探索時間が膨大になる．Ishiguro et al は教示によりこの問題を回避している．また Asada et al [12] はゴール状態近傍からタスクを遂行できる状態を超楕円体で近似しながら離散化する手法を提案した．彼らはサッカーロボットによるシュート行動を実機で獲得させた． Ueno et al [13] は Asada et al らの手法をオンライン化したアルゴリズムを提案した．彼らは簡単なナビゲーションに適用し，コンピュータによるシミュレーションで検証した．しかしこれらの手法は Asada et al の指摘もあるように適切な状態空間を獲得するためには偏りの無い十分なデータが必要である．また状態を超楕円体で近似するため，凹状の状態への対応が困難である．†††

本論文ではロボットが自身の経験に基づいて逐次的にセンサー空間を分割して現実的な時間内に目的の行動を学習する手法を提案する．すなわち，センサ出力に基づくセンサ変化と報酬信号の関数近似による局所モデルを構築し，このモデルに基づいてセンサ空間を分割することによって状態空間を構成する．この手法は以下の特徴を持つ．

(1) 学習初期から局所モデルに基づいて行動するのでゴール到達が早い時期から可能．また，局所モデルが有効な限り新たに状態を作らないので無駄な探索過程が削減でき，学習の収束が速い．
(2) 逐次学習であるので環境の変動に対応できる．
(3) 状態の形を規定していないので凹状の状態への対応可能である．

本手法をボールをゴールにシュートするサッカーロボットに適用した．計算機によるシミュレーションと実機による結果を示す．実機のロボットは一時間半以内に状態空間を逐次的に分割してシューティング行動を獲得した．

以下ではまず簡単に強化学習について触れ，提案する手法の基本的な考え方とアルゴリズムを示す．次に適用したタスクを説明した後，計算機シミュレーションと実機システムでの実験結果を示す．

2. 強化学習の基本的枠組み

ロボットが識別することができる状態の集合を $\mathbf{S}$ とし，環境に対してとり得る行動の集合を $\mathbf{A}$ とする．環境は現在の状態とロボットの行動によって確率的に遷移するマルコフ過程に従うものとする．状態と行動の組 (s, a) に対しては報酬 $r(s, a)$ が定義される．

一般的な強化学習の問題は，減衰する報酬の積算を最大にする政策を見つけることである．環境のダイナミクスを学習しながら政策を決定するものとして， Watkins の Q 学習アルゴリズム [1] は有効な手法である．

Q 学習では，状態 $s \in \mathbf{S}$ において行動 $a \in \mathbf{A}$ をとり，次状態 s' に遷移した時，行動価値関数値 $Q(s, a)$ を以下のように更新する．

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

ここで， α は学習率， γ は減衰係数である． Q 値が与えられると，各状態 s に対して $Q(s, a)$ が最大となる行動 a を選ぶことによって政策が定義される．

以下で述べる本研究では s をオンラインの行動学習過程でどのように構成するかという問題を扱っている．

3. 基本的な考え方

強化学習を行うのに適した状態空間は「行動を切り替える必要のある状態が適切に分割されている」必要がある．環境やロボット，与えられるべきタスクがあらかじめわかっていないときに適した状態空間を学習を始める前に設計者が用意しておくことは困難である．

タスクを行うための適当な状態空間をロボット自身が自分の経験を通して構成するためには「行動を切り替える必要がある状態（状況を）発見する」能力が必要である．しかし「行動を切り替えるべき状況」は与えられるタスクに依存するので，事前に発見する指標を獲得するのは難しい．

そこで本手法ではこの「行動を切り替えるべき状態」の候補を挙げるための指標として「センサ出力に基づくセンサ信号の変化と報酬信号の関数近似による局所モデルの近似度」を採用する．ロボットは各種の行動のそれぞれに対して逐次局所モデルを構築し，この指標に基づき状態空間を分割することで，各種の行動に意味のある領域を発見する．

「行動を切り替えるべき状態」を発見する指標を逐次構築される局所モデルから得ることによって以下の利点が生まれる．

- タスク遂行のための行動シーケンスをあらかじめ知っておく必要がない．
- 未経験領域や分割必要のないところは必要以上に分割しないので状態数は少なくてすみ，その結果無駄な探索過程を削減でき強化学習が速やかに収束する．
- 分割過程の逐次性により環境の変化に対応できる

4. アルゴリズム

Fig.1に提案する手法の大まかな流れを示す．まず，ロボットはセンサー出力をデータとして取得する．データは現在の局所モデルと比較され，誤差が許容範囲内であれば局所モデルを新しいデータで更新する．そうでなければロボットは新しい局所モデルを作成し状態空間を構成し直し，今までの経験から得られた知識を再利用することで行動価値関数を初期化し強化学習する．

†††粗く近似した場合誤差が大きくなり，細かな凸状態で近似すると状態数が膨大になる．

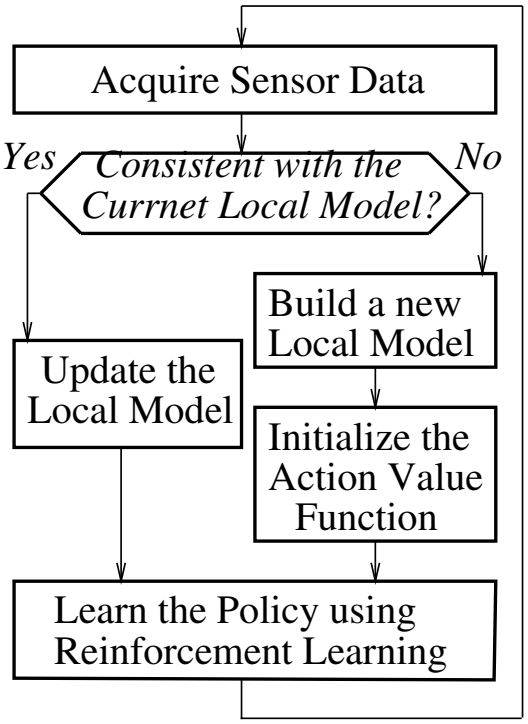


Fig. 1 The brief flow of the proposed method

4.1 行動の定義

従来の強化学習においては「行動」とは「固定された一定時間のモーターコマンドの実行」と定義されていた．現実の環境ではこの定義では Asada et al が指摘しているように「状態と行動のずれ問題」が生じる [14]．彼らはロボットの前進や回転といった単位時間あたりの基本的な動きを基本動作と呼び，「現在の状態が変化するまでの基本動作のシーケンス」を行動と再定義することでこの問題に対処した．ここでは彼らの行動の定義に従う．

4.2 データ構造

センサ出力を $x_0, x_1, \cdots$，報酬を r とする．本手法では報酬もセンサ変化の一部と考えて

$$\boldsymbol{x} = (x_0, x_1, \cdots)^T \tag{2}$$

$$\dot{\boldsymbol{x}}' = (\dot{x_0}, \dot{x_1}, \cdots, r)^T \tag{3}$$

と記述する．基本動作を $m_i \in M\ (i = 1, \cdots, n_m)$ とするとデータ $d \in D$ は次のように定義される．

$$d =< m_i, \boldsymbol{x}, \dot{\boldsymbol{x}}' > \tag{4}$$

すなわちセンサ出力 $\boldsymbol{x}$ で動作 m_i をとった時のセンサ出力変化および報酬 $\dot{\boldsymbol{x}}'$ の三つ組である．ある状態が構成された時のデータ d の $\boldsymbol{x}$ をその状態におけるセンサ情報の代表点と呼ぶ．新たなセンサ出力が入って来た時，このセンサ出力に一番近い代表点が所属する状態をその状態と分類する．この分類により局所モデルの誤差が大きければ新たな状態を構成し，それに対する局所モデルを構築する．

代表点を設ける理由は以下の通りである．すべてのデータを記録することはデータ量が時間に比例して増加していくので現実的でない．またセンサのノイズや環境の変動，モータコマンドの不確実性により同じセンサ出力同じ基本動作でもセンサ出力がばらつき，正しいデータが取れない場合がある．そこで十分に近いデータに対しては $\dot{\boldsymbol{x}}'$ の更新をするにとどめることでこれらの問題に対処する．また複数の代表点で一状態の領域を表現するので，領域が凹状であっても対処できる．

4.3 局所モデルの構築

まずはじめに局所モデルを構築する方法を説明する．局所モデルとしてセンサ出力の勾配の線形モデルを利用する．

$$\dot{\boldsymbol{x}}' = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \tag{5}$$

局所モデルの構築と分割のアルゴリズムは以下の通り．

- 基本動作 $m_i \in M\ (i = 1, \cdots, n_m)$ について
  (1) $C :=$ 動作が m_i の全ての $d \in D$
  (2) C を用いて式 (5) を最小自乗近似する．
  (3) もし残差の不変分散が大きければ
    (a) C を重み付ユークリッドノルムを類似度したクラスター分析の手法を用いて 2 つに (C_1 と C_2) 分割する
    (b) $C := C_1$．(2) に進む．
    (c) $C := C_2$．(2) に進む．

ゴール状態付近ではセンサ出力に基づくセンサ変化の関数近似により分割された領域が適切であるとは限らない．なぜなら同じ分割された領域で同じ行動をとっても同じ領域に遷移しない場合 (例えばゴール状態到達の成功失敗) がよくあるからである．しかし式 (3) において $\dot{\boldsymbol{x}}'$ に報酬を含めているので報酬信号によってもセンサー空間を分割することになり，これがゴール状態付近での領域分割に役立っている．

4.4 状態空間の構築

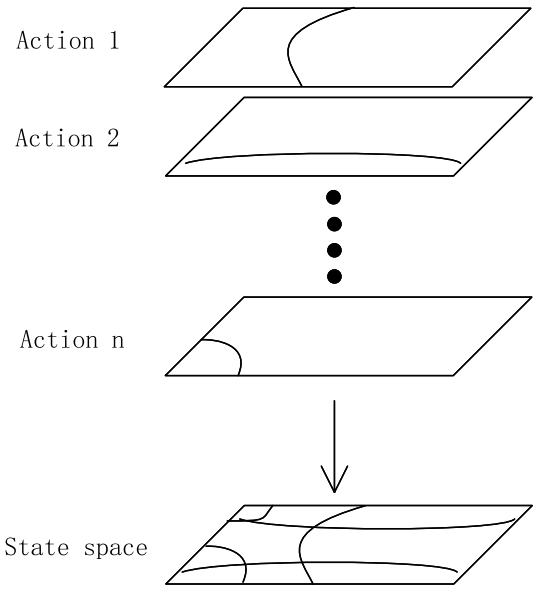


Fig. 2 The construction of state space

4.3節では各種の行動で局所モデルを逐次構築していくことを述べたが，ここではそれらを使って状態空間を構成する方法を述べる． Fig.2にその例を示す．各種の行動に対して局所モデルの構築により，センサ空間が分割されている．分割された領域は各種の行動にとって特徴のある領域と考えられる．しかしながら行動の種類によって状態空間が異なっている場合，強化学習を直接適用できない．そこで各種の行動によって分割されたセンサ空間を重ねあわせる (論理積) ことによって状態空間を構成する．

4.5 経験から得た知識の再利用

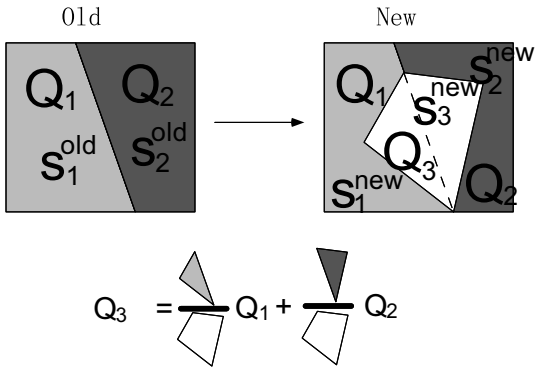


Fig. 3 Recalculation of Q value

理論的には新しい状態空間が構成されるたびごとに行動価値関数を初期化するべきである．しかしそれでは学習によって過去の経験から得た知識を有効に利用できない．そこで古い状態空間とその行動価値関数から新しい状態空間の行動価値関数を計算することで知識を再利用することを考える．

ここでは古い行動価値関数の重みつき和で新しい行動価値関数を計算する．重みは新しくできた状態の張る空間に含まれる古い状態の空間の割合とする．割合は新しい状態と古い状態両方に含まれるセンサー出力の代表点の数によって計算される． Fig.3に概念図を示す．

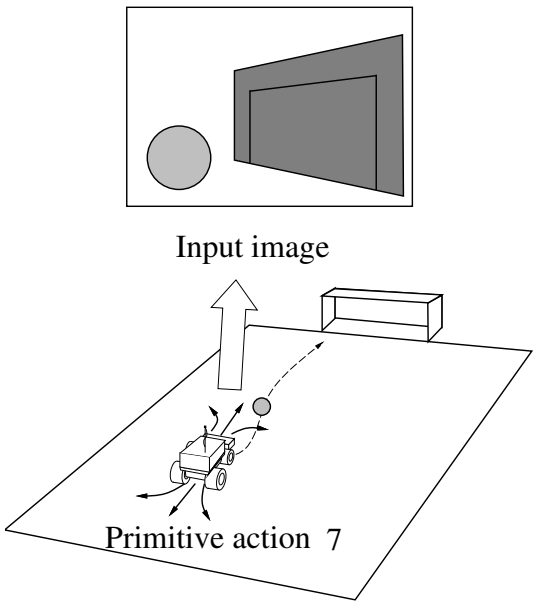
5. タ ス ク

本研究で扱うタスクは Fig.4(a) に示すようにボールをゴールにシュートすることである．

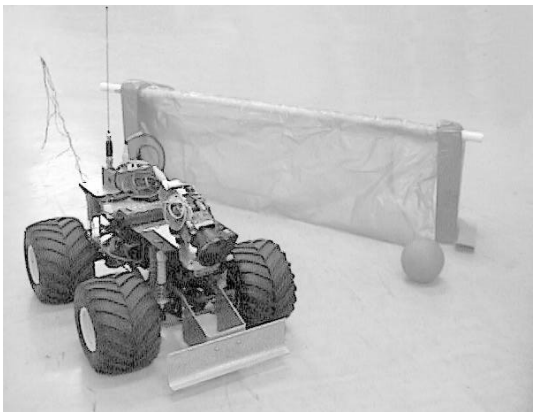
環境にはボールとゴールがあり，ロボットはそれに備え付けられたテレビカメラからボールとゴールの基本的な画像特徴を得ることができる．ロボットはボールやゴールの大きさや距離などの三次元情報、カメラパラメータ、ロボット自身の動特性などの先験的な知識は一切与えられていない． Fig.4(b) に実際に用いた移動ロボット，ボール，ゴールを示す．

ロボットが識別できるセンサー情報はボールとゴールの状態を表わす５次元ベクトルである．ボールについては画面上における大きさと位置，ゴールについては大きさと位置に加え，さらに傾きである．

カメラの画角 (65°) が狭いため，ボールやゴールを見失いがちである．この時ボールやゴールを表現する特徴量がなくなるが，一時刻前の状態を記憶しておくことでボールやゴールがどちらの方向に見失ったか知ることができる．このような見失った状態で



(a) The task is to shoot a ball into the goal



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

Fig. 4 A task and our real robot

は絶対値の大きい正負の符号がついた一定の値をそれらの特徴量として与える． Fig.5 はセンサー出力が１次元の時の局所モデルの構成とセンサー空間の分割の例を示している．結果としてこのような状態の局所モデルは傾きがゼロである状態として得られる．

6. 実 験 結 果

6.1 シミュレーション

ロボットの基本動作は Fig.4(a) に示すように左前進，前進，右前進，左後退，後退，右後退，停止のあわせて７通りである．一試行は①ロボットがシュートに成功する，②ロボットもしくはボールがフィールドの外に出る，③ロボットがゴールポストに衝

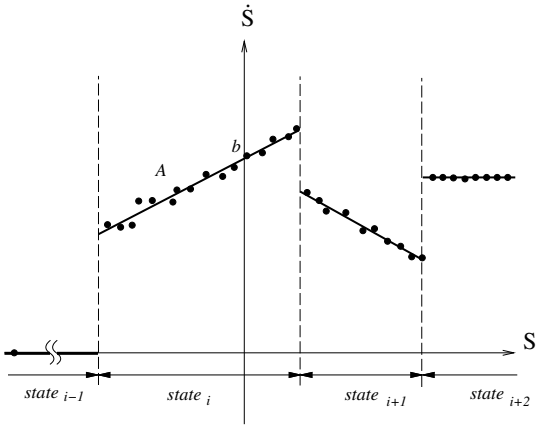


Fig. 5 The construction of local model and the segmentation of sensor space

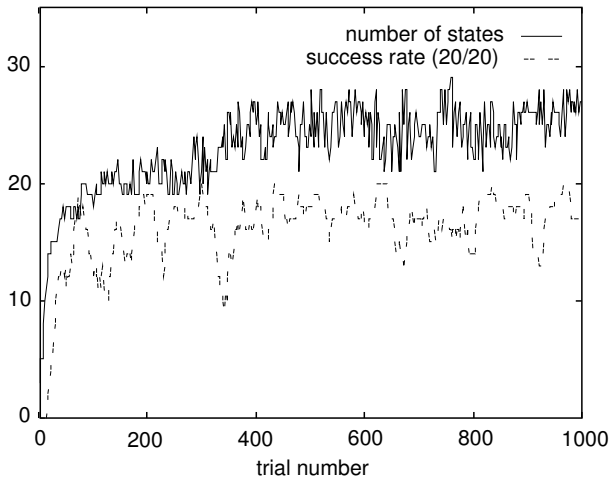


Fig. 6 The success rate and the number of states

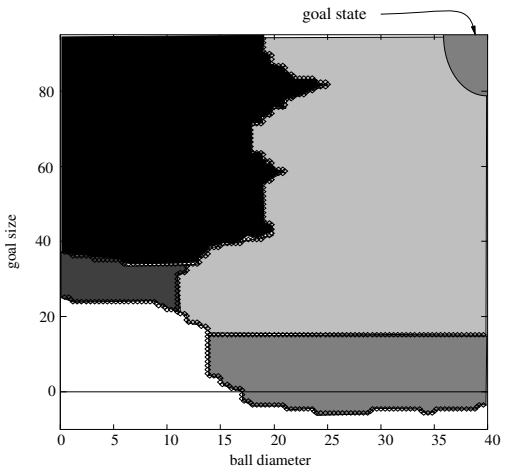


Fig. 7 Result of state space construction

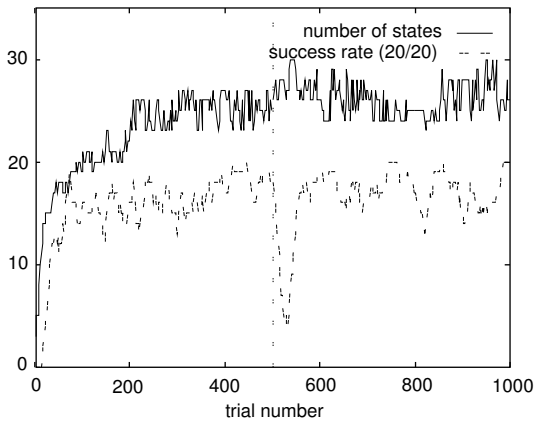


Fig. 8 Teh success rate and the number of states in the case that environment change one the way

突する，のいずれかである．試行が終わると，ロボット，ボールをリセットする．報酬としてはロボットがボールをゴールに入れた時 1.0 を与え，それ以外は 0.0 とした．ロボットは 90% の確率で各時点での最適行動をとり，10% の確率でランダム行動を取る．

Fig.6に本手法を適用したときの成功率と分割された状態の数を示す．ただし成功率は最近の 20 試行の間に成功した回数である．

Fig.7は試行回数 1,100 で最終的に得られた状態空間のボールとゴールが同時に見える時の様子を示す．ただし，5 次元の状態空間に対し，ボールの位置は画面中央，ゴールは傾きが無く真ん中に見えている状態での断面をとり，ボールの大きさとゴールの大きさの 2 次元で表現したものである．

また環境が動的に変化しても対応できることを示すために試行回数 500 でボールの大きさを倍に変化させた時の成功率と状態数を示す図を Fig.8に示す．一端成功率が下がるが，過去の知識を利用して即座に適応している様子が伺われる．

6.2 実機での実験

Fig.9に実際に構築したシステムを示す． Fig.10(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) とボールおよ

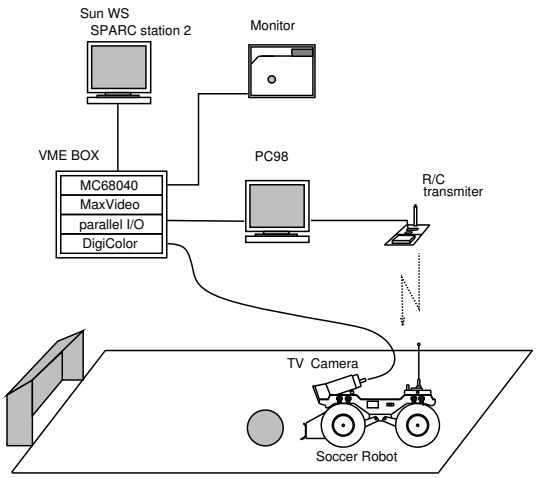
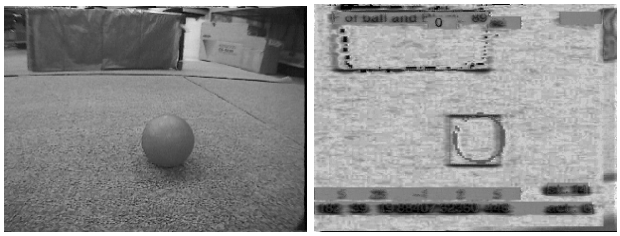


Fig. 9 A configuration of the real robot



(a) input image (b) detected image

Fig. 10 Detection of the ball and the goal

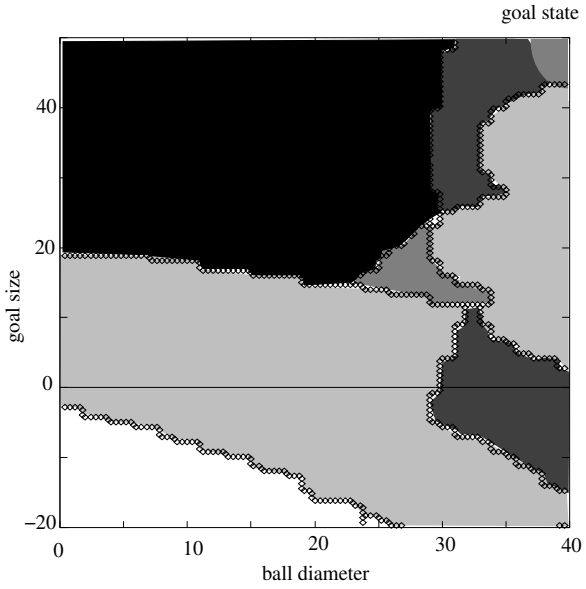


Fig. 11 state space construction of real robot experiment

びゴールを実時間パイプラインビデオ画像プロセッサ Datacube MaxVideo 200 で検出した画像を示す．状態識別，行動選択は Vxworks OS が搭載されているホスト CPU(MC68040 CPU) 上で行なわれ，この CPU はイーサネットで Sun のワークステーションと接続している．サンプリング時間が約 30[ms] である．基本動作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．(詳細は文献 [15] 参照)．

Fig.11は試行回数 72 までで得られている状態空間のボールとゴールが同時に見える時の様子を示す (見方は Fig.7に同じ)．またこの時得られた状態数は 18 であった．

Fig.12はロボットがシューティングに成功した例を，6 つの時刻に分けて示している．センサノイズによる状態の誤認識やモータコマンド出力の不確実性のために何度か試行錯誤を繰り返しながらタスクを遂行する．①でロボットはボールをゴールにシュートするために前進している．しかし②でロボット自身の速度が速すぎ，曲がりきれずにボールを蹴ることに失敗している．②でボールはその時ロボットに隠れている．それでロボットは③で

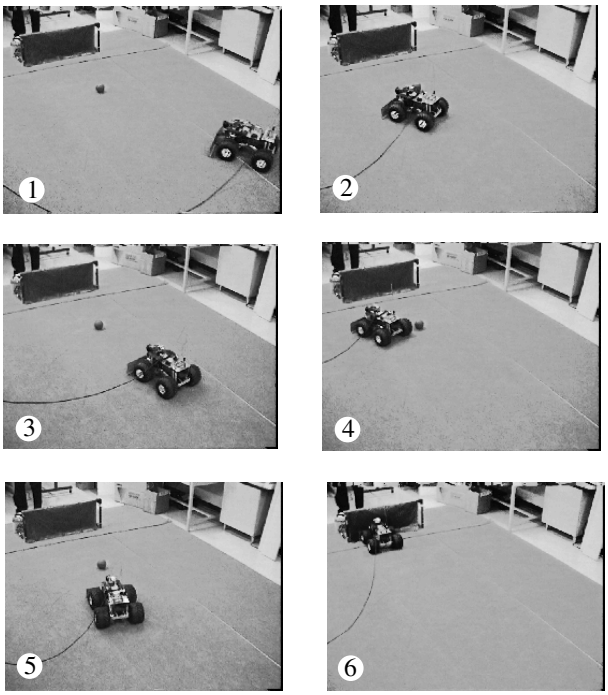


Fig. 12 The robot succeeded in shooting a ball into the goal

ボールをシュートできるように左にバックしている．しかし④で再び失敗しているが，⑤でまた左後退をしている．最終的にロボットは⑥でシュートすることを成功している．

7. おわりに

本論文ではロボットの経験を基にセンサー空間を逐次的に分割していく手法を提案し，ロボットが実時間内に目的の行動を学習することを示した．

今後の課題として，より複雑な関係の理解を必要とするタスクやセンサ情報の選択の問題へ拡張が挙げられる．

参考文献

[1] C. J. C. H. Watins and P. Dayan. “Technical note: Q-learning”. *Machine Learning*, Vol. 8, pp. 279–292, 1992.

[2] Jonalthan H. Connell and Sridhar Mahadevan. *ROBOT LEARNING*. Kluwer Academic Publishers, 1993.

[3] Steven D. Whitehead and Dana H. Ballard. Active Perception and Reinforcement Learning. In *MLWS*, pp. 179–188, 1990.

[4] Justin Boyan and Andrew Moore. Generalization in Reinforcement Learning: Safely Approximating the Value Function. In *Proceedings of Neural Information Processings Systems 7*. Morgan Kaufmann, January 1995.

[5] Richard S. Sutton. Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding. In *Advances in Neural Information Processing Systems 8*, pp. 1038–1044, 1996.

[6] Fuminori Saito and Toshio Fukuda. Two-Link-Robot Brachiation with Connectionist Q-Learning. In *Proceedings of the third international conference on simulation of adaptive behavior (From animals to animats 3)*, pp. 309–314. The MIT Press, 1994.

[7] J. S. Albus. A New Approach to Manipulator Control: The Cerebbellar Model Articulation Controller (CMAC). *Journal of*

Dynamic Systems, Measurement, and Control, Trans. ASME, Vol. 97, No. 3, pp. 220–227, Sept 1975.

[8] J. S. Albus. Data Storage in the Cerebellar Model Articulation Controller (CMAC). *Journal of Dynamic Systems, Measurement, and Control, Trans. ASME*, Vol. 97, No. 3, pp. 227–233, Sept 1975.

[9] Ben J. A. Kröse and Joris W. M. van Dam. Adaptive state space quantisation for reinforcement learning of collision-free navigation. In *Proceedings of the 1992 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vol. 2, pp. 1327–1331, 1992.

[10] Artur Dubrawski and Patrick Reignier. Learning to Categorize Perceptual Space of a Mobile Robot Using Fuzzy-ART Neural Network. In *Proceedings of the IEEE/RSJ/GI International Conference on Intelligent Robots and Systems*, Vol. 2, pp. 1272–1277, September 1994.

[11] Hiroshi Ishiguro, Ritsuko Sato, and Toru Ishida. Robot Oriented State Space Construction. In *Proceedings of the 1996 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vol. 3, pp. 1496–1501, 1996.

[12] 浅田稔, 野田彰一, 細田耕. ロボットの行動獲得のための状態空間の自律的構成. 日本ロボット学会誌, Vol. 15, No. 6, pp. 886–892, 1997.

[13] A. Ueno, K. Hori, and S. Nakasuka. Simultaneous Learning of Situation Classification Based on Rewards and Behavior Selection Based on the Situation. In *Proceedings of the 1996 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vol. 3, pp. 1510–1524, 1996.

[14] 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.

[15] 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, pp. 73–78, 1994. Q 学習のロボットビジョンシステム.

高橋 泰岳

1972 年 12 月 13 日生. 1994 年大阪大学大学院学研究科博士前期課程修了. 現在, 同年同大学博士後期課程在学中. 知能ロボットの行動獲得に関する研究に従事.

浅田 稔

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.