論　文

Actor-Qアーキテクチャに基づく能動認識学習システム**

柴田　克成†　　西野　哲生††*　　岡部　洋一††

Active Perception Learning System Based on Actor-Q Architecture

Katsunari SHIBATA†, Tetsuo NISHINO††*, and Yoichi OKABE††

あらまし　Q-learningとActor-Criticアーキテクチャの Actorを組み合わせたActor-Qアーキテクチャとそれを用いた能動認識学習システムを提案する．Actor-Qアーキテクチャは，システムの出力を，離散的な意思である「行動」と連続値である「動作」に区別し，まず，Q値を用いて「行動」を決定し，その「行動」が「動作」を伴う場合には，更に，該当するActorの出力に従って「動作」を行う．そして，Q値は，Q-learningで学習し，Actorは，その「行動」に対応するQ値をCriticの出力として学習を行う．ここでは，センサの移動及び各パターンであるという認識の結論を下すことを行動とし，それぞれにQ値を割り当てる．そして，センサの移動が選択された際は，Actorの出力に従ってセンサを移動する．認識が選択された場合は，対応するパターンであるという認識結果を出力し，正解不正解を表す強化信号によって該当するQ値を学習する．Q値計算部及びActorはともにニューラルネットで構成し，視覚センサ信号を入力とする．これによって，従来の能動認識学習システムの問題点，(1) 認識に対する評価値の極大値にセンサがトラップされること，(2) 各時間ステップごとに認識出力を評価し，連続値の強化信号を与える必要があること，(3) 認識結果を出力するタイミングをシステム自身で判断できないの三つを解決することができる．そして，不均一なセンサセルをもつ視覚センサを用いたシミュレーションによって上記の効果を確認した．

キーワード　強化学習，ニューラルネット，能動認識，Actor-Qアーキテクチャ，視点移動

1. まえがき

人間をはじめとする生物は，その行動を行う上で必要とされる外界の情報を，様々な感覚器より獲得している．しかし，外界の情報は膨大であり，すべての情報を細部まで獲得することは困難，かつ，大変効率が悪いと考えられる．これに対し，生物は能動的に感覚器を動かす，いわゆる能動認識の機能を有し，効率的に，必要な情報を獲得している．

人間の感覚器の中で，得られる情報量が最も大きい視覚センサ（目）を見てみると，網膜上のセンサセルの分布は不均一であり，センサ全体を使って外界の様子を大まかにとらえ，中心部のセンサセルの分布が密な部分を適切な場所に移動して詳細を把握するという役割分担によって，効率的に正しい認識を行っていると考えられる．我々人間は，未知のものでも必要に応じて学習し，認識できるようになる．よって，認識対象がどのようなものであり，それを認識するためにどこを見るべきかといった認識対象に関する知識は，生得的にもっているのではなく，生後の学習によって獲得されていると考えられる．

我々が工学的にパターン認識を行う場合には，前処理としてセグメンテーションを行って，パターンの切出しを行い，切り出されたパターンから特徴を抽出し，その特徴から認識を行うという手法が一般的である[1],[2]．また，福島らは，ニューラルネットを用いて，位置ずれの吸収と特徴抽出を繰り返すことにより，より生物に近い形で位置ずれに強いパターン認識を実現している[3]．

また，近年，「脳」の立場からも，「ロボット」の立場からも，認識における能動性が注目されてきている[4],[5]．前者では，選択的注意のモデル[6]をベー

† 大分大学工学部電気電子工学科，大分市
Dept. of Electrical and Electronic Engineering, Faculty of Engineering, Oita University, 700 Dannoharu, Oita-shi, 870–1192 Japan
†† 東京大学先端科学技術研究センター，東京都
Research Center for Advanced Science and Technology, The University of Tokyo, 5-3-1 Komaba, Meguro-ku, Tokyo, 153–0041 Japan
* 現在，(株)日本オラクル
** 本研究における基本的なアイデアは，1997年3月の本会ニューロコンピューティング研究会にて発表したものである．

スにしており，主に「注意」という形で，パターンのセグメンテーションを行うことに重点が置かれている．また，視覚センサ信号をリカレントニューラルネットへの入力として，認識の学習をさせるだけで，過去の視覚センサ信号から文脈を抽出し，それを保持し，次の認識に対して選択的注意を行うことができるようになること，更には，その文脈の記憶の際に，連想記憶の能力があることも示されている [7]．一方，ロボットの分野では，実際に，視覚センサを移動させるシステムが構築されている [5] が，移動物体の発見や追跡に主眼が置かれており，認識を目的として，その注視点を学習するというものではない．

更に，近年，「強化学習」が，その自律性，適応性，合目的性の観点から注目を集めている．従来，強化学習は，主に，行動計画のための学習としてとらえられてきた．しかし，ニューラルネットにセンサ信号を直接入力し，出力信号を用いてモータを動かし，強化学習を適用することによって，センサからモータまでの，認識や注意なども含めた総合的な機能を調和的に学習させるという方向も示されている [8]．このような中で，Whitehead らは，ブロックの積み替え問題において，着目するブロックの選択を強化学習の一つである Q-learning [9] によって獲得させている [10]．しかしながら，認識を陽に扱ったシステムにはなっておらず，また，着目するブロックを選択しても，そこにセンサを動かすための動作は考慮されていない．

これに対し，筆者らは，強化学習を適用することによって人間のように適切なセンサの動作を学習によって獲得し，効率的なパターン識別を行う工学的なシステムの構築とともに，生物のセンサ動作の獲得に強化学習が用いられている可能性を示すことを目指してきた．そして，簡単な視覚センサ信号そのものをニューラルネットへ入力し，認識結果に対する評価のみから，連続値としてのセンサの動作を学習する方法を提案してきた [11]．しかしながら，この方法では，遅延報酬に対応したシステムとなっていなかったため，毎単位時間ごとに認識結果を出力し，そのときの教師パターンと出力パターンの差を連続値スカラの評価値としてもらわなければならなかった．そして，その結果，認識出力のパターンが，局所的に教師パターンに近いところでトラップされ，結果的に正解が得られない場合があるという問題点があった．また更に，システムが最終的に認識結果を下すタイミングを，ある一定時間経過後としていたため，認識できる位置まで移動しても，時間がくるまで認識結果が出せなかったり，逆に，認識できる位置に移動する前に時間がきて，正しい認識結果を出力できないという問題点があった．また，この時間を決めるためには，あらかじめ与えられる認識問題を予測しなければならず，強化学習の自律性，柔軟性が十分に生かされたとはいえなかった．

本論文では，はじめに，システムの出力を，離散的な「行動」と連続値としての「動作」に分けて，出力と学習を行う Actor-Q アーキテクチャを提案する．そして，それを能動認識学習システムに適用することによって，センサを移動させるか認識結果を出力するかを決定する機能を付加し，センサの移動とともに，強化学習によって獲得することを提案する．そして，不均一センサセル密度をもつ視覚センサを用いて，パターンの中で注視すべき点にセンサセルの密度が高い部分を移動させ，自ら，認識結果を出力するタイミングを判断し，正しくパターンを認識することができるかどうかを確認する．

2. Actor-Q アーキテクチャ

本章では，本論文で提案する Actor-Q アーキテクチャとそれに基づく能動認識学習システムについて説明する．まず最初に，システムの出力を離散的な意思である「行動」と，連続値出力である「動作」の二つに区別する．システムは，始めに「行動」を決定し，その行動が「動作」を必要とする場合に，更に，「動作」を決定する．例えば，「走る」「歩く」「止まる」という「行動」の中から「歩く」が選択されたときに，実際に「歩く」ために足の筋肉への指令をいくつにするかという「動作」を決定する．行動決定には Q 値を用い，その学習には Q-learning を用いる．一方，「動作」の決定には Actor-Critic [12] の Actor を用いる．通常，Actor の学習には Critic の出力が用いられるが，Critic の学習も，Q-learning も，ともに TD (Temporal Difference) 学習 [13] を基本としていることから，別途 Critic を設けることなく，その「行動」に該当する Q 値を Critic の出力として用いる．これを Actor-Q アーキテクチャと呼ぶ．

そして，これを図 1 のように，Q-ネットと Actor-ネットの二つのニューラルネットで構成し，Q-ネットの各出力は各行動に対応させ，Actor-ネットの各出力は動作信号の各出力先に対応させる．また，入力が同じであれば，入力層と中間層を共有して，二つのネットワークを一つで構成することも可能である．更に，

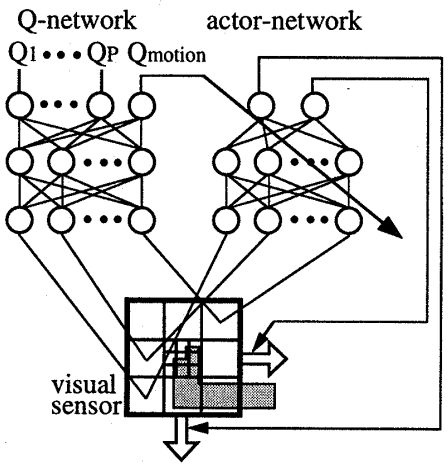


図1 Actor-Q アーキテクチャに基づく能動認識学習システム

Fig. 1 The active perception learning system based on Actor-Q architecture proposed in this paper.

複数の「行動」が「動作」を必要とする場合は，複数のActor-ネットを有し，選択された行動によってActor-ネットのゲーティング，つまり，選択された行動に該当するActor-ネットの出力を選択的に「動作」として出力することも可能である．

次に，能動認識学習システムへのActor-Q アーキテクチャの適用について述べる．ここでは，“パターン p であるという認識結果を出力する”こと及び“センサを移動させる”ことをそれぞれ行動としてとらえる．提示パターンが P 個ある場合は，図1のように，個々のパターンであるという認識の結論を下す P 個の行動と，センサ移動の1個を併せた $(P+1)$ 個の行動があり，それぞれに対し Q 値を割り当てる．そして，Q 値に基づいて一つの行動が選択される．

選択された行動が「認識」の場合には，それが正解であるか不正解であるか，つまり，正解ならば報酬がもらえ，不正解ならば何も与えられずに一つの試行が終了する．ただし，試行終了後の Q 値は0.0とするので，パターン p であるという結論を下す「認識」の行動“$Recog(p)$”に対するQ-learningの式は，

$$Q(s(t),''Recog(p)'') = (1-\alpha)Q(s(t),''Recog(p)'') + \alpha r \qquad (1)$$

となる．ただし，$s(t)$：時刻 t でのセンサ入力（状態），α：学習係数，r：報酬であり，正解のときに1.0，不正解のときに0.0とする．これは，サルに認識の学習をさせる際に，正解すると報酬がもらえるという状況に似ている．ここでは，ニューラルネットを学習させるので，

$$Q_{train}(s(t),''Recog(p)'') = r \qquad (2)$$

を教師信号として，繰り返さずに，1回だけ，そして，該当する Q 値だけを学習させる．

一方，「センサ移動」の行動が選択された際には，Actor-ネットの出力に従ってセンサを移動させる．この場合は，試行は終了せず，センサ移動後に新しい入力を得て，再び次の行動が選択される．Q 値は，報酬が0の場合の通常のQ-learningに従って，

$$Q_{train}(s(t),''motion'') = \gamma \max_a Q(s(t+1),a) \qquad (3)$$

という教師信号によって，やはり1回だけ，該当する Q 値のみを学習させる．ここで，γ：割引率，a：とり得る行動とする．

センサの移動 $\mathbf{m}$ は，ここでは，x 軸，y 軸方向のセンサ速度を規定するものとし，Actor-ネットの出力 $\mathbf{o}_m$ に乱数 $\mathbf{rnd}$ を加え，

$$\mathbf{m} = \mathbf{A}(\mathbf{o}_m + \mathbf{rnd}) \qquad (4)$$

とする．ただし，$\mathbf{A}$:対角成分のみの定数行列とする．そして，Actor-ネットは，

$$\mathbf{O}_{m,train} = \mathbf{O}_m + (\gamma \max_a Q(s(t+1),a) - \max_a Q(s(t),a))\mathbf{rnd} \qquad (5)$$

という教師信号によって学習する．この式は，通常のTD学習を行うActor-Criticにおける，強化信号が0の場合のActorの学習式，

$$\mathbf{O}_{m,train} = \mathbf{O}_m + (\gamma P(x(t+1)) - P(x(t)))\mathbf{rnd} \qquad (6)$$

の $P(x(t))$ を $\max_a Q(s(t),a)$ に置き換える，つまり，その状態での各行動の評価値の中で最も良いものをCriticの出力（その状態の評価値）とすることによって導くことができる．ただし，$P(x(t))$ は時刻 t でのCriticの出力（状態評価値）である．もし，行動選択がgreedy，つまり，Q 値の最も大きい行動が選択されるようになれば，式(5)の最後の $\max_a Q(s(t),a)$ は $Q_{motion}(s(t))$ となる．

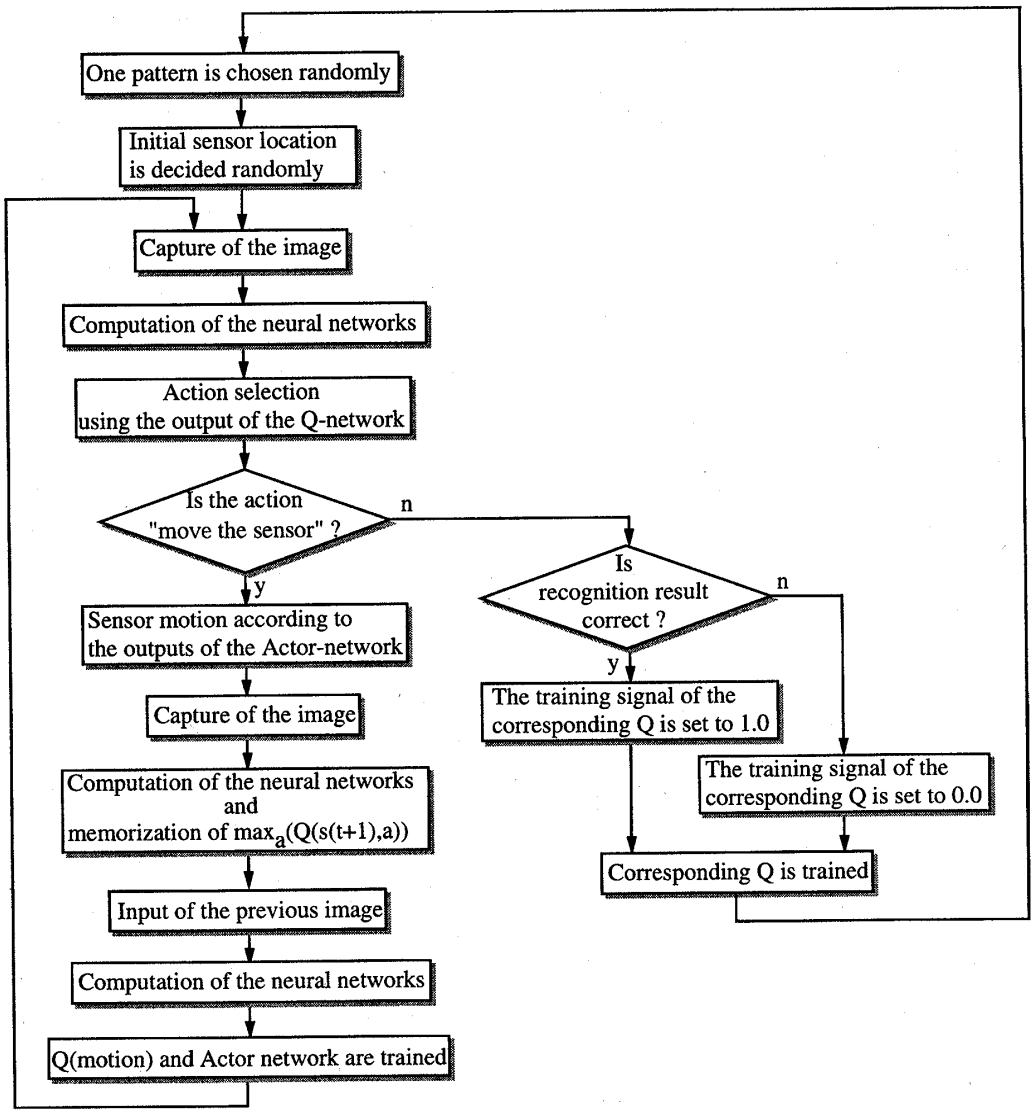


図2　学習の流れを示すフローチャート
Fig. 2　The flow chart of the proposed learning.

学習は，Q-ネット，Actor-ネットともにBP（Back Propagation）法[14]によって行う．以上の流れを示したフローチャートを図2に示す．

Q-learningとActor-Criticを統合したアーキテクチャは，Morimotoらによっても提案されている[15]．しかし，彼らは，高次元空間の学習において，サブゴールを設定し，Q値によって，現在の状態でどのサブゴールを選ぶかを決定し，そのサブゴール達成をActor-Criticで学習させている．したがって，目的としては，高次元空間の分割を目的としている点で異なり，手法としては，Q値をCriticの信号に置き換えるということはしておらず，各Actor-CriticにCriticが存在しているという点で異なる．

3. 視点移動のシミュレーション

3.1 問題設定と学習

本論文では，ある複数のパターンを用意し，提示されているものがどのパターンであるかを識別する問題を扱う．用いるセンサは，図3のように，2次元の視覚センサで，センサの中心部の解像度は大きく，周辺部は解像度が小さいという不均一性を導入した．

このセンサに対し，認識すべきパターンを提示する．しかし，パターンの提示位置は毎試行ごとに可変であり，またその提示位置を表すような情報も与えられない．パターンの初期提示位置は，システムの視界に全くパターンが入らない場合は除外するが，図4(a)のように，パターンの一部のみが視界にとらえられてい

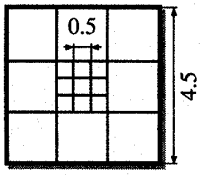


図 3　本論文で用いた不均一なセンサセルをもつ視覚センサ
Fig. 3　Visual sensor with non-uniform sensory cells employed in this paper.

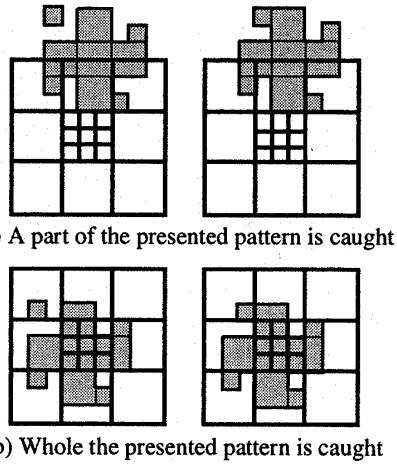


図 4　パターンの識別ができない場合
Fig. 4　Two cases in which the presented pattern cannot be identified.

るような場合も含まれる．また，パターン全体が視覚センサの中に入っていても，図 4 (b) のように，パターンを識別できない場合もある．このような場合は，センサ入力から識別を行うことは不可能であり，センサを適切な場所に移動する必要がある．図 4 (a) のような場合には，パターンの重心方向に視覚センサを移動させればよいが，図 4 (b) のような場合には，識別すべきパターンによって適切な移動方向が変化する．システムは，センサを移動させ，最終的にどのパターンであるかの結論を下し，それが正解かどうかが強化信号として与えられる．そして，学習によって適切なセンサ移動と認識の結論を下すタイミングを獲得させる．

3.2　タスク設定

図 5 に，本論文で用いた 2 種類のパターンセットを示す．センサは前述の図 3 のように，1 辺が 4.5 × 4.5 の 2 次元視覚センサで，17 個のセンサセルよりなる．中心部の小さなセルは 9 個あり，1 辺 0.5，周辺部の大きなセルは 8 個あり，1 辺 1.5 である．個々のセンサ信号は，投射されたパターンが各センサセルの受容野中に占める割合とし，ニューラルネットに入力するときは，−1.0 から 1.0 に線形変換した値とする．セン

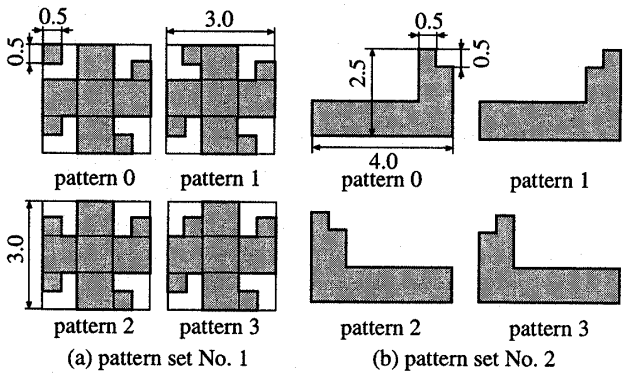


図 5　提示したパターンセット
Fig. 5　The presented pattern sets.

サの初期位置は，すべてのセンサセルからの信号の総和が 0.5 以上の範囲で試行ごとにランダムに決定される．パターンセット 1 が提示された場合は，パターンを識別するためには，左上の領域に視覚センサを移動する必要があり，パターンセット 2 の場合は，まず全体を見て，パターン 0 または 1 なのか，パターン 2 または 3 なのかを判断し，更にそのいずれかを区別するために，前者の場合は右の方へ，後者の場合には左の方へセンサを移動させる必要がある．

ニューラルネットは，3 層とし，Q-ネットの中間層ニューロン数は 30 個，Actor-ネットの中間層ニューロン数は 10 個とした．各ニューロンの出力関数は，値域が −0.5 から 0.5 のシグモイド関数 $(1/(1+\exp(-x))-0.5)$ とし，式 (3) の Q 値の教師信号は，実際には，

$$O_{train} = \alpha(Q_{train} - 0.5) \tag{7}$$

と変換して用いた．ただし，α:定数である．また，逆に，ニューロンの出力 O は，

$$Q = O/\alpha + 0.5. \tag{8}$$

と変換して Q 値とした．ここでは，ニューロンの出力関数の飽和領域を避けるため α は 0.8 とし，出力が −0.4 から 0.4 の間に入るようにした．もし，Q 値が 0.0 以下になった場合は 0.0 とした．また，学習前には，入力にかかわらず出力が一定値となるように，中間層から出力層への初期重み値をすべて 0.0 とした．更に，学習が不安定になったり，一様なセンサ速度が簡単に出現しないように，出力層ニューロンのバイアスを，Actor-ネットでは 0.0，Q-ネットでは −2.2 と固定し，学習させなかった．Q-ネットでのバイアス値を −2.2 とすると，出力がほぼ −0.4，つまり，Q 値

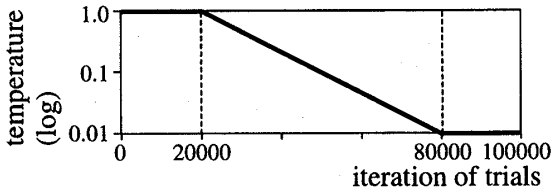


図 6　行動選択で用いるボルツマン選択の温度変化
Fig. 6　Temperature cooling schedule that is used in the action selection.

の初期値がどのような状態でもほぼ 0.0 になる．

また，学習のための試行回数は 100,000 回とし，式 (3) の Q 値の学習に用いる割引率 γ は 0.99，式 (4) の Actor の出力からセンサの移動量への変換を決める定数 β は 0.4 とした．したがって，1 単位時間の最大移動距離は，x, y 方向ともに 0.2 となる．また，乱数 **rnd** は -1.0 から 1.0 の範囲の三つの一様乱数を掛け算したものとした．行動選択には，学習中はボルツマン選択を用い，そのときの温度は，図 6 のように 1.0 から 0.01 まで少しずつ下げていった．また，学習後は，Q 値が最大の行動を選択し（greedy policy），センサ移動も，乱数を付加しなかった．各試行は，パターンがセンサから消えた場合も含めて，システムが認識の結論を出すまで続けた．しかし，学習初期では，行動選択の温度が高く，ランダムに近いため，センサ移動が選択され続けることはなかった．また，学習が進んでも，センサからパターンが消えた状態で，センサ移動が選択され続けると，式 (3) の Q 値の学習によって Q 値の値がどんどん小さくなるため，センサ移動の選択回数の上限を定めなくても，無限にセンサ移動が選択されることはなかった．

3.3 結　　果

学習後に実際に認識を行ったところ，いずれのパターンセットでも，センサ移動後に「認識」の行動が選択され，正しく認識を行うことができるようになった．パターンセット 1 の学習後に，0.25 間隔の 132 の初期センサ位置からのセンサ移動の様子を図 7 に示す．いずれのパターンを提示しても，また，センサの初期位置をどこにしても，センサの中心がパターンの左上に移動している様子がわかる．また，最終的なセンサ位置は，パターン 0，1，2 の場合にはほぼ 1 点に収束し，中心部の小さなセンサセルの一つが，ちょうど各パターン間の差異となっている小さな正方形をとらえていることがわかる．また，パターン 3 の場合は，収束する場所は他に比べて広くなっている．この傾向は，ニューラルネットの初期重み値を変化させても同じであった．

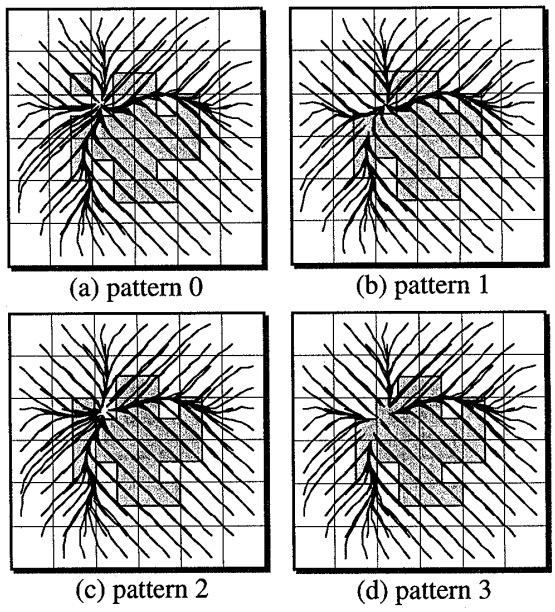


図 7　パターンセット 1 が提示されたときの視覚センサの軌跡
Fig. 7　The trajectories of the visual sensor when the pattern set No.1 was presented.

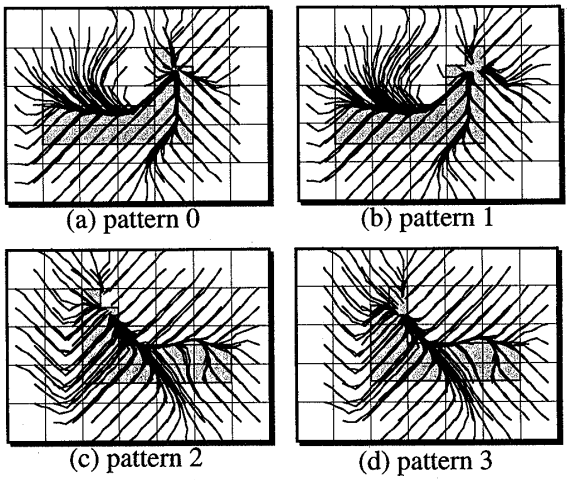


図 8　パターンセット 2 が提示されたときの視覚センサの軌跡
Fig. 8　The trajectories of the visual sensor when the pattern set No.2 was presented.

図 8 にパターンセット 2 の場合の視覚センサの軌跡を示す．パターン 0 と 1 の場合には，センサをパターンの右上に，パターン 2 と 3 の場合には左上に移動してから認識結果を出力しており，提示パターンによってセンサの移動方向が異なっていることがわかる．また，図 9 には，各パターンを提示し，センサ位置を変化させたときの，提示パターンに対応する Q 値の分

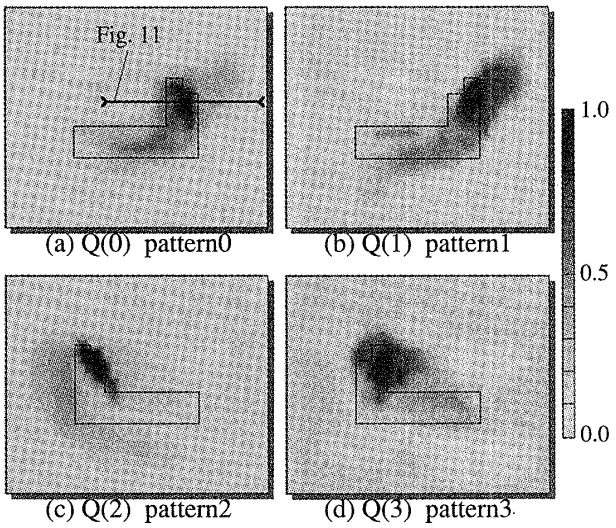


図 9　提示パターンに対応する Q 値の分布

Fig. 9　The distribution of the Q-values that is corresponding to the presented pattern.

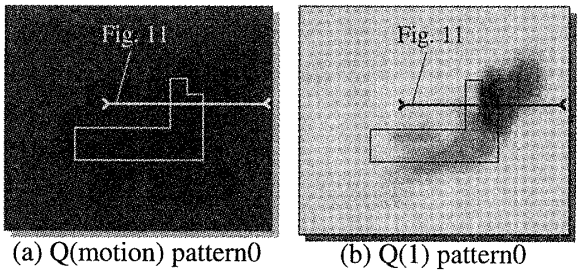


図 10　パターン 0 が提示されたときのセンサ移動の Q 値とパターン 1 の Q 値の分布

Fig. 10　The distribution of the Q-values for the sensor motion and the Q-value for the pattern 1 when the pattern 0 was presented.

布を示す．やはり，パターン 0，1 の場合には，パターンの右上にセンサがあるとき，パターン 2，3 の場合には，左上にあるときに Q 値が大きくなっていることがわかる．また，パターン 0 が提示されたときの，センサ移動の Q 値とパターン 1 に対応する Q 値の分布を図 10 に示す．センサ移動の Q 値は，センサ位置によらず全体的に大きな値を示している．また，パターン 1 の Q 値は，パターン 0 の Q 値と同様に，センサがパターンの右上にあるときに大きな値となっていることがわかる．

これらの図では，各行動に対する Q 値の相対的な大小関係がわかりにくいため，図 9 (a) 及び図 10 (a) (b) の中の横線で示した断面で Q 値の分布を見た図を図 11 に示す．このことから，矢印で示した認識可能な部分では，いずれの Q 値も大きな値となっているが，中でもパターン 0 の認識に対する Q 値が最も大

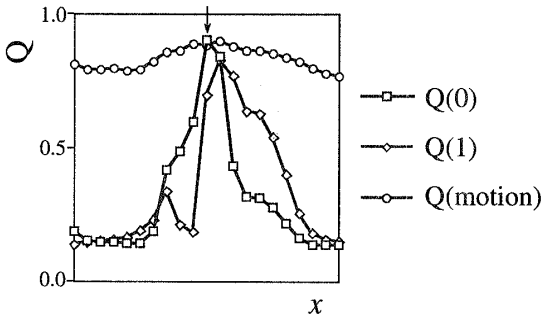


図 11　Q 値の分布曲面の断面

Fig. 11　The one-dimension of the Q-value distribution when the sections of the Q-value surfaces were observed.

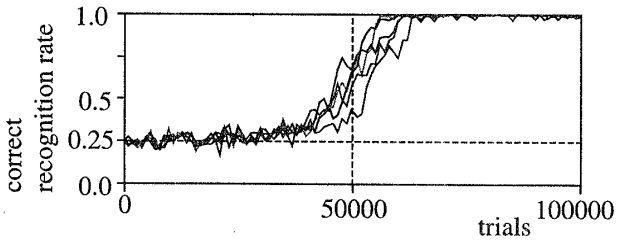


図 12　学習曲線．y 軸は正しい認識結果を下した割合

Fig. 12　Learning curve when the pattern set No.2 was presented. The y axis indicates the probability of the successful recognition.

きな値となっていることがわかる．また，それ以外のところでは，センサ移動用の Q 値が最も大きな値となっており，パターン 1 の Q 値が一番大きな値になって誤認識するところはなかった．センサ移動の Q 値はなだらかに下がっているが，この傾斜は，式 (3) の Q 値の学習に用いる割引率 γ の値による．これを 1 に近い値とすることにより，センサ移動の Q 値の傾斜がなだらかになり，認識がうまくできないところで認識の Q 値に局所的なピークがあっても，センサ移動が選択されて，そこにトラップされないことがわかる．

ニューラルネットの初期重み値を変えて 5 回のシミュレーションを行って，学習の経過とともに認識率がどのように変化するかを図 12 にプロットした．縦軸は，正しく認識できた確率を表している．5 回とも，50000 試行ぐらいで認識率が急に大きくなり，ほぼ同じような学習曲線を描いていることがわかる．図 13 に，50000 試行学習後のセンサの軌跡を示す．この図より，センサはある程度移動することができるようになっているが，識別できない領域でも認識の Q 値が大きくなって，結論を下していることがわかる．そして，結果として，パターン 1 を提示しているにもかかわらず，システムはパターン 0 であるという間違った

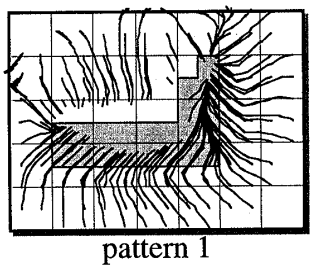


図 13　50000 試行学習後にパターン 1 が提示されたときの軌跡

Fig. 13　The trajectories of the visual sensor when the pattern 1 was presented after 50000 trials of learning.

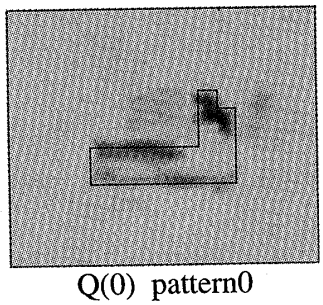


図 14　センサを移動させないで，認識の Q 値だけ学習した場合の Q 値の分布

Fig. 14　The distribution of Q-value when the sensor did not move and only the Q-values for recognition were trained.

結論を導いていた．

次に，センサ移動の行動を選択しないで，認識の Q 値のみを学習させたときの Q 値の分布を図 14 に示す．この図を，図 9 (a) と比較すると，パターン右上のピークが小さく，逆に，パターンの左上部の山が大きくなっていることがわかる．このことより，センサを移動させないと，Q 値の学習がうまく進まないことがわかる．この原因を考察してみる．提示パターンの区別ができないところでは，同じ入力に対し，違う教師信号が与えられる．センサを移動させない場合は，ほとんどそのような場所で学習しなければならないため，その影響で，正しく認識できる場所でもうまく学習が進まないと考えられる．センサが移動するようになると，移動しているところでは認識の Q 値は学習されないため，学習される領域が限定されてくる．すると，更に正しく認識できるところでは，該当する認識の Q 値がより大きくなり，それによって更にセンサの移動の学習が進み，相互に影響を及ぼし合いながら認識率が向上していくと考えられる．

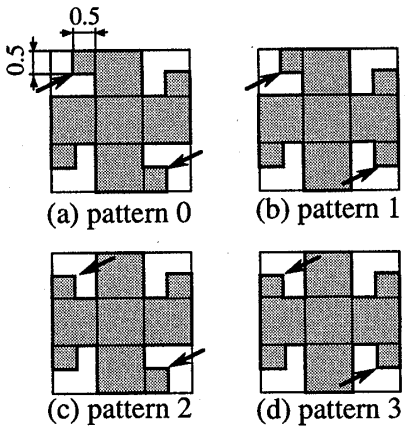


図 15　パターン識別のために文脈入力が必要なパターンセット

Fig. 15　The pattern set in which the system requires the context inputs to identify each presented pattern.

3.4　文脈によるセンサ移動

我々が文字やパターンを認識する際には，文脈から次に何が来る可能性があるかを予測することで，非常に効率的に認識を行うことができる．この機能を実現するためには，文脈の抽出と，その文脈をいかに使うかが問題となる．ここでは，その準備段階として，文脈が既に抽出されたと仮定し，抽出された文脈を利用したセンサ移動ができるかどうかを確認する．そこで，図 15 のように，パターンの 1 箇所にセンサの中心をもっていっても，どのパターンか特定できないようなパターンセットを用意する．この場合，左上にセンサが移動しても，パターン 0 か 1 か，または，パターン 2 か 3 か区別できない．また，右下に移動しても，パターン 0 か 2 か，またはパターン 1 か 3 か区別できない．そして，各パターンに対応する四つの入力を追加し，現在提示されている可能性のあるパターンに 1 の信号を，そうでないものに 0 の信号を与える．この場合には，パターン 0 か 1 とわかっていれば，右下に，パターン 0 か 2 だとわかっていれば，左上にセンサを動かせばよい．ここでは，文脈入力が 1 になる個数は常に 2 個とし，0 と 1，0 と 2，1 と 3，2 と 3 の 4 通りのうちのいずれかとした．

この学習は，前述の学習に比べて大変難しく，Q-ネットの中間層を 50 個，Actor-ネットの中間層を 20 個として，2000000 試行の学習を行った．また，式 (3) 中の γ は 0.96 とした．行動選択の温度変化は図 6 と同様とし，x 軸を拡大した．学習後のセンサ移動の様子を図 16 に示す．この図より，文脈によって，センサ

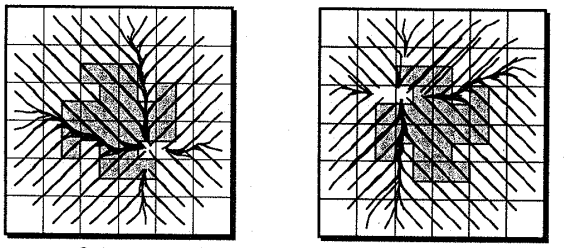


図 16　文脈入力によるセンサ軌道の差異
Fig. 16　The difference of the sensor trajectories depending on the context inputs.

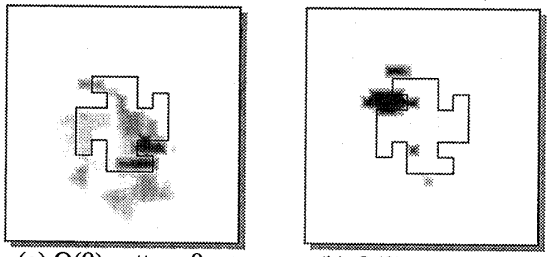


図 17　文脈入力による Q 値の分布の差異
Fig. 17　The difference in the Q-value distribution depending on the context inputs.

の軌道が変化していることがわかる．また，このときの，Q 値の分布を図 17 に示す．Q 値のピークの場所が文脈情報によって大きく異なっていることがわかる．

このタスクでは，同じパターンが提示されていても，文脈情報によってセンサの移動方向を変化させなければならない．ところが，γ を 0.99 とすると，パターン 0 が提示されて，文脈情報が 0 か 1 を示しているときのセンサ移動を先に学習すると，文脈情報が 0 か 2 のときもセンサが右下に移動してしまう．そして，センサ移動の Q 値が広い範囲で大きくなるため，パターンの左上部で「認識」を行って，認識の Q 値を学習することが少なくなり，うまく学習が進まないという状況となった．この，パラメータ設定の難しさと学習の遅さの解決は，今後の課題である．

3.5 考　　察

本節では，工学的なパターン認識システムとして，及び，生物の視覚のモデルとしての妥当性に関して考察する．パターン認識システムにおいては，一般的に，パターンの平行移動だけでなく，拡大・縮小や回転などにも対応できることが望ましい．本システムでは，パターンの平行移動のみ考慮しており，拡大・縮小や回転は考慮していない．拡大・縮小や回転を学習させることを考えた場合，基本的なアルゴリズムとしては，本システムと同一のものが適用できると考えられる．ただし，その場合，動作の自由度が増えるため，その分学習速度が問題となる．また，より現実に近い問題に適用するという意味でも，システムのハードウェア化や事前知識の導入など，学習速度に対する対策が必要であると考えられる．

一方，本研究は，生物の視覚のモデルとしての側面ももつ．しかし，図 7，図 8 や図 16 などの視覚センサの軌跡を見ると，始点から終点まで直線的に向かわず，x 方向 y 方向それぞれ等速度で進んでいる部分が多く，眼球の動作として明らかに不自然である．これは，センサ動作が x 方向 y 方向それぞれの速度で規定されており，ダイナミックスが導入されておらず，更に，速度の最大値が規定されているためであると考えられる．また，マニピュレータのリーチング運動に強化学習を適用した際に，ダイナミックスを導入しても学習できる [16], [17] ことから，これらの不十分な点は視覚システムの問題であり，アーキテクチャの問題ではないと考えられる．報酬のみから視覚センサの動作を学習できたことから，生物の視覚システムにおいても，眼球の動作を強化学習によって獲得している可能性を示すことはできたのではないかと考える．

最後に，本論文で提案した Actor-Q アーキテクチャは，能動認識の学習のみならず，離散的な行動選択を要する問題一般に有効であると考えられる．例えば，移動ロボットが障害物を発見した際に，その右側を通り抜けるか左側を通り抜けるかといった問題において，単一の Actor-Critic アーキテクチャで学習させると，時として障害物の正面で立ち止まってしまったり，障害物にぶつかってしまうという問題がある [8]．しかし，Actor-Q アーキテクチャを用いて，右に行くか左に行くかを先に決めることができれば，これらの問題は回避できると予想される．また，複数の行動目標があった場合に，どちらの目標を目指すべきかを決定することも可能であると考えられる．ただし，Q-learning では，とり得る行動をあらかじめ設定しなければならないが，強化学習の自律性，適応性を生かすためには，必要に応じて行動を割り当てられるような仕組みが必要であると考えられる．

4. む　す　び

システムの出力を離散的な行動と連続値の動作に分け，各行動に対して Q 値を割り当て Q-learning で学習し，動作の方は Actor-Critic の Actor の出力とし

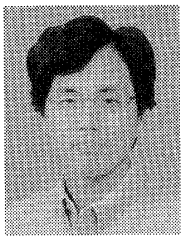
て，該当する行動の Q 値を Critic の代わりとして学習する Actor-Q アーキテクチャを提案した．そして，Actor-Q アーキテクチャを能動認識学習システムに適用することによって，(1) 最終的な認識出力を行うタイミングをシステム自身が判断し，(2) その後に与えられる正解か不正解かを示す強化信号のみから学習し，(3) 認識の評価値の局所的なピークにとらわれず，認識できる場所までセンサを移動し，正しい認識を行うことを提案した．そして，不均一なセンサセルを有する視覚センサを移動させてパターンを認識するタスクを行い，その能力を確認した．また，文脈入力を設けることにより，センサ入力が同一の場合でも，文脈によって異なったセンサ動作を実現できることを示した．

謝辞 本研究の一部は，文部省科研費重点領域「創発システム」(No.264) の補助のもとで行われました．ここに謝意を表します．

文　献

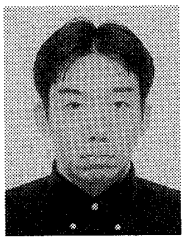
[1] 鳥脇純一郎，パターン情報処理の基礎，朝倉書店，1998.
[2] 大津展之，栗田多喜夫，関田　巌，パターン認識 理論と応用，朝倉書店，1996.
[3] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift position," Biol. Cybern., vol.36, no.4, pp.193–202, 1980.
[4] 福島邦彦，"能動的視覚情報処理神経回路モデルによる研究，" 文部省科学研究費重点領域 "高次脳機能のシステム的理解，" 平成 10 年度報告書，1999.
[5] S. Rougeaux and Y. Kuniyoshi, "Robust real-time tracking on an active vision head," Proc. IEEE-RSJ Int'l Conf. of Intelligent Robots and Sys. (IROS)'97, Grenoble, Sept. 1997.
[6] K. Fukushima, "A neural network for visual pattern recognition," IEEE Comput., vol.21, no.3, pp.65–75, 1988.
[7] 柴田克成，伊藤宏司，"認識の学習に基づく注意と連想記憶の形成，" 信学技報，NC99-137, March 2000.
[8] 柴田克成，岡部洋一，伊藤宏司，"ニューラルネットワークを用いた Direct-Vision-Based 強化学習，" 計測自動制御学会論文集，vol.37, no.2, pp.168–177, Feb. 2001
[9] C.J.C.H. Watkins and P. Dayan, "Q-learning," Machine Learning, vol.8, pp.279–292, 1992.
[10] S.D. Whitehead and D.H. Ballard, "Learning to perceive and act by trial and error," Machine Learning, vol.7, pp.45–83, 1991
[11] 柴田克成，西野哲生，岡部洋一，"強化学習による能動認識能力の学習，" 日本神経回路学会誌，vol.3, no.4, 126–134, 1996.
[12] A.G. Barto, "Adaptive critics and the basal ganglia," Models of Information Processing in Basal Ganglia, pp.215–232, MIT Press, Cambridge, MA, 1995.
[13] R.S. Sutton, "Learning to predict by the methods of temporal differences," Machine Learning, vol.3, 1988, pp.9–44, 1988.
[14] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning representations by back-propagating errors," Nature, vol.323, no.9, pp.533–536, 1986.
[15] J. Morimoto and K. Doya, "Hierarchical reinforcement learning of low-dimensional subgoals and high-dimensional trajectories," Proc. ICONIP'98, vol.2, pp.850–853, Oct. 1998.
[16] K. Shibata, M. Sugisaka, and K. Ito, "Hand reaching movement acquired through reinforcement learning," Proc. The 2000 KACC (Korea Automatic Control Conference), 90rd (CD-ROM, 4pages), Oct. 2000.
[17] 柴田克成，杉坂政典，伊藤宏司，"強化学習によるリーチング動作の獲得，" 信学技報，NC2000-170, March 2001.

（平成 12 年 10 月 2 日受付，13 年 1 月 29 日再受付）



柴田　克成　（正員）

1989 東大大学院工学系研究科機械工学専攻修士課程了．1989 (株) 日立製作所入社．1992 年 10 月同社退職．1993 東大大学院工学系研究科先端学際工学専攻博士課程中退．1993 東大先端科学技術研究センター助手．1997 東工大大学院総合理工学研究科リサーチアソシエイト（日本学術振興会未来開拓学術研究推進プロジェクト研究員）．2000 大分大学工学部電気電子工学科講師．現在，同助教授．主として，ニューラルネットを用いた強化学習・自律学習システムの研究に従事．博士（工学）．



西野　哲生

1997 東大大学院工学系研究科情報工学専攻修士課程了．同年，(株) 日本オラクルに入社．

岡部　洋一　（正員）

1972 東大大学院工学系研究科電子工学専攻博士課程了．1972 東大工学部電気工学科講師，同助教授，IBM San Jose 研究所客員研究員，教育用計算機センター助教授，電子工学科助教授，電子工学科教授を経て，現在，同大学先端科学技術研究センター教授．主として，高速高機能デバイス，特に，超伝導，脳磁計測，ニューラルネットの研究に従事．工博．