[計測自動制御学会論文集
Vol.37, No.2, 1/10 (2001)]

ニューラルネットワークを用いた Direct-Vision-Based 強化学習 †

―センサからモータまで―

柴田克成*・岡部洋一**・伊藤宏司***

Direct-Vision-Based Reinforcement Learning Using a Layered Neural Network

- For the Whole Process from Sensors to Motors -

Katsunari SHIBATA*, Yoichi OKABE** and Koji ITO***

In this paper, Direct-Vision-Based Reinforcement Learning is proposed not only for the learning of motion but for the learning of the whole process, which includes recognition, from sensors to motors in robots. In this learning, raw visual sensory signals are put into a layered neural network directly, and the network is trained by Back Propagation using the training signal generated based on reinforcement learning. By employing neural network, whole the process becomes seamless and is trained purposively and harmonically.

Two simulations of the mobile robot with visual sensors are performed as examples. One task requires stereo-vision, and the other requires obstacle avoidance. By observing the hidden representation in the neural network, it is shown that some abstract representation of spatial recognition is formed without any advance knowledge.

Finally it is shown that each visual sensory cell makes a role of localization of the global information of the space and it helps the fast and stable learning.

Key Words: reinforcement learning, neural network, visual sensor, hidden representation, localization

1. はじめに

近年，ゲーム[1]やプランニングの問題[2]への適用を始め，強化学習が注目を集めている．これは，強化学習が，報酬や罰といった少ない情報から行動を学習するため，学習の自律性・柔軟性が非常に高いということが一番の理由であると考えられる．この学習の自律性・柔軟性は，外界へ動作として出力し，それによって変化する外界の状態を入力として受けるというフィードバックループを利用することによって実現される．この意味で，センサとモータを有するロボットはその典型であり，強化学習が最も威力を発揮する適用先の一つであると言える．

ロボットに用いられるセンサの中で，視覚センサは，多数のセンサセルを有し，外界に対する多くの情報をロボットに提供する．また，われわれ人間を振り返ってみても，外界の情報の獲得に対する視覚センサ (目) への依存度が非常に大きいことが容易にわかる．ロボットで視覚センサを扱う場合，まず画像処理を行って，必要な情報を抽出し，それを利用してロボットの行動を決定するというのが一般的な流れである．

浅田らは，視覚センサを有するサッカーロボットへ強化学習を適用し，さらに，実機でその機能を確認するという先駆的な研究を行っている．彼らは，ボールやゴールの大きさや見える位置を，予め用意したプログラムによって視覚センサ信号から計算させ，それによってできる空間をいくつかの状態に分割し，各状態に対する行動を Q-learning によって学習させている[3][4]．また，この場合，いかにして状態空間を構成するかが問題となるが，ゴールまでの状態遷移と動作系列を見て自律的に状態を分割する方法も提案している[5]．

ここで，視覚センサ信号を前処理して，状態空間を構成する理由を考える．一般に視覚センサは多数のセンサセルより構成されるため，信号数が多い．したがって，個々のセンサ信号の違いを別の状態として学習させると状態数が膨大となり，学習に非常に時間がかかると予想される．さらに，個々の視覚センサセルの受容野が局所的であるため，汎化もきかないということが理由として挙げられる．また，収束性という面から見ると，視覚センサ信号を連続値としてそのまま扱うと，離散状態とならず，マルコフ決定過程の下で収束性が

† 第４回創発システムシンポジウムにて発表 (1998・8)
* 大分大学工学部電気電子工学科，大分市大字旦野原 700 番地
** 東京大学先端科学技術研究センター，東京都目黒区駒場 4-6-1
*** 東京工業大学大学院総合理工学研究科知能システム科学専攻，横浜市緑区長津田町 4259
* Dept. of Electric & Electronics Engineering, Fuculty of Engineering, Oita Univ., 700 Dannoharu, Oita
** Research Center for Advanced Science and Technology, Univ. of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo
*** Dept. of Comp. Intel. and Sys. Sci., Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama
(Received June 6, 2000)

証明されている Q-learning の収束性が保証されなくなることも理由として考えられる．さらに，われわれ人間においても，動作を行う際に個々のセンサ信号が意識下にのぼらないことから，視覚信号から大域的な空間情報への変換は予め与えられ，その大域的な情報の上で動作が学習されているように見えることも一つの理由と考えられる．

しかしながら，前処理をするためには，予めタスクに関する知識が必要である上，前処理という設計者の手が加わることにより，強化学習の自律性・柔軟性といった最大の利点を阻害する可能性がある．たとえば，前述のサッカーロボットの場合，タスクの中で，ボールやゴールが出てくること，そして，視覚センサ上でその大きさや位置が変化すること，さらに，それがタスクの遂行に意味を持つことを知らなければ設計できない．また，環境が変化する場合には，その前処理も適応的に変化できることが望ましいと考えられる．

従来，強化学習は，運動系，特にプランニングの学習としてとらえられてきている．しかし，より良い行動を行うためには，適切な認識，注意，記憶などタスクによって様々な機能が要求される．したがって，強化学習が潜在的にそれらを学習する能力を持ちあわせていると考えられる．筆者らの一部は，視覚センサ信号をニューラルネットへの入力として，認識やセンサ移動を学習できること[6]，簡単な視覚センサ信号から移動ロボットの動作を直接学習できることを示した[7]．また，前述の Tesauro の研究[1]でも，視覚センサ信号ではないものの，ゲームの盤面の情報を直接ニューラルネットに入力し，学習させている．

本論文では，まず，単に，視覚センサ信号を直接ニューラルネットへ入力しても学習できるというだけでなく，それが，ロボットがセンサからモータまでの過程を合目的的，調和的に学習し，知識を獲得するのに必要であること，そして，強化学習＋ニューラルネットの枠組みが，単なる組み合わせではなく，互いにその欠点を補い合い，ロボットの自律学習にとって質的な変革をもたらすことを主張する．そして，移動ロボットが，両眼立体視や障害物回避を必要とするタスクを学習できることをシミュレーションで示すとともに，単に強化学習を行うだけで，ニューラルネットの中間層において，必要に応じた知識 (ここでは空間認識) の獲得と抽象化が起こることを示す．最後に，センサ信号は無数にあるため，直接入力すると学習が遅くなるとの議論，および，ニューラルネットと強化学習を組み合わせると学習が不安定になるとの指摘[8]に対し，センサ信号の局所性が，学習を高速化，安定化し，それらの問題を回避する働きをすることを示す．

2. 強化学習とニューラルネットの組み合わせ

本章では，ロボットの知能化において，強化学習とニューラルネットの組み合わせが有効な理由を説明する．

2.1 強化学習から見た有用性

前述のように，従来の一般的な強化学習は，状態空間を形成し，極端に記述すると，各状態と行動のテーブルを学習によって形成していると言える．しかし，この場合，自律的な学習ではあるものの，学習の結果を他のタスクに利用することは難しく，知能と呼ぶには単純すぎると考えられる．これに対し，システム全体を強化学習で合目的的に学習させることを考える．ここでもし，認識部，プランニング部，制御部などと機能別にモジュール化させると，そのインタフェースをあらかじめ規定しなければならず，その分柔軟性に欠ける．さらに，行動の試行錯誤による結果を他のモジュールの学習にどのように反映させるかを決定することは困難である．逆に，各モジュールごとに試行錯誤をすることはたいへん効率が悪い．そこで，ニューラルネットでシステム全体をシームレスに構成し，強化学習に基づいて内部生成された教師信号をもとに，バックプロパゲーション (BP) 法[9]でそのニューラルネット全体を学習させれば，システム全体が調和的，合目的的，かつ，動作の試行錯誤だけで効率的に学習できると考えられる．また，ニューラルネットの中間層に，空間認識などの知識が必要に応じて蓄えられ，その中間層を共有することで，その知識を他のタスクでも利用することが容易であると考えられる．さらに，ニューラルネットは，連続値の非線形関数近似が可能である上，リカレント構造にすることによって，記憶やダイナミクスの形成も可能になると期待される．

2.2 ニューラルネットから見た有用性

ニューラルネットの学習として，前述の BP 法がある．BP 法は教師あり学習としてたいへん有用であるが，教師信号が必要であり，それ自体は自分で学習していく能力は持っていない．結局，人間が教師信号という形で知識を与えることが要求され，プログラミングによって知識を与えて知能化することに対する差異があいまいになり，学習の利点を十分に発揮できなかった．

それに対し，強化学習は，外界とのフィードバックループを用いて自律的に学習するアルゴリズムである．したがって，そのアルゴリズムに基づいてニューラルネットの教師信号を内部生成して学習させれば，システムの外側から見ると，自律的に学習していることになり，従来の “人間による知識の付与” から “ロボットによる知識の獲得” へという大きな飛躍に結び付くと期待される．

3. 時間軸スムージング学習に基づく強化学習

ロボットの学習には，一般に Q-learning が良く用いられる．しかし，Q-learning では行動の離散化が必要であり，離散化のためには，事前知識が必要となる．さらに，ロボットの動作は一般に連続値で表現でき，そのほうが滑らかな動作を実現できる．よって，ここでは actor-critic アーキテクチャを採用する．ただし，actor(動作生成部) と critic(状態評価部) は入力が同じであるため，同一のニューラルネットで構成し，中間層を共有する．また，critic の学習には，時間軸スムージング (TS, Temporal Smoothing) 学習[7]を採用した．はじめに，時間軸スムージング学習について簡単に説明する．

3.1　時間軸スムージング学習

本学習は，空間情報を，その情報が出現する時間の情報にマッピングするための学習であり，空間情報（センサ信号）をニューラルネットに入力し，その出力の時間の2階微分値を0に近づけるように学習する．つまり，時間軸に対する出力曲線の凹凸をなくすという学習を行う．たとえば，視覚センサ信号を入力と考え，センサの前を物体が滑らかに動いている場合，各センサセルは局所的な受容野しか持たないが，発火が時間的に近いセンサセルは空間的にも近い位置にあると考えられる．したがって，時間軸スムージング学習で学習すれば，その出力は，時間を表すだけでなく，物体の位置を表すことにもなる[11]．このようにして，時間軸スムージング学習は，空間情報が時間の経過に対して滑らかに変化しているという前提の下で，時間を媒介として，空間の情報を抽出する能力を有している．詳細は[7)11)]を参照されたい．

3.2　時間軸スムージング学習に基づく強化学習

ここでは，強化学習を，与えられた一つの目的を最短時間で達成するための動作系列を経験から学習するものとする．目的達成までの所要時間による評価を獲得するため，状態評価値の学習には，前述の時間軸スムージング学習を適用する．しかし，これだけでは，各試行間で初期状態が異なると，評価値の時間変化量を必ずしも同一にすることはできない．つまり，評価値からだけでは目的達成までの所要時間を試行間で平等に評価することができない．そこで，評価値の時間変化量が経路によって変化しないように，時間軸スムージング学習を式(1)に置き換える．まず，状態評価値の教師信号 p_s を，一つ前の時間の状態評価値 $p(t-1)$ を用いて，

$$p_s(t-1) = p(t) - \frac{P_{max} - P_{min}}{N_{max}[i]} \qquad (1)$$

ここで，P_{max}, P_{min}：評価値の理想値域の最大値，最小値，ここでは，$P_{max} = 0.4$, $P_{min} = -0.4$

とし，Back Propagation(BP)法[9)]で学習する．$N_{max}[i]$は，試行 i の時点での目的達成までの最大所要時間を表す．これによって，評価値の時間変化量は $(P_{max} - P_{min})/N_{max}[i]$ に近づき，時間による2階微分値も0に近づいていく．$N_{max}[i]$ は，試行 i の際の所要時間 $N[i]$ を用いて，

$$N_{max}[i] = \begin{cases} N[i] & if N[i] > \lambda N_{max}[i-1] \\ \lambda N_{max}[i-1] & otherwise \end{cases} \qquad (2)$$

ここで，$\lambda = 1 - 1/\tau$, τ：大きい時定数

と計算し，問題の複雑さと学習の進行状況に応じて N_{max} を適応的に変化させる．これによって，一試行あたりの目的達成までの所要時間が大きいと，評価値の時間変化量は小さくなり，学習の進行とともに所要時間が減少してくると，評価値の時間変化量が大きくなって，評価値の値域を常に有効に使うことができる．ここでは，ニューラルネットの出力関数は入力層以外すべてシグモイド関数とし，値域は-0.5から0.5とした．そして，ロボットが目的を達成したときに教師信号を0.4として学習させる．

一般的に，criticの学習には，TD(Temporal Difference)学習が用いられる．この場合，目指すべき目標が1つであり，報酬の大きさを1とすると，

$$\bar{r} = \gamma^T \qquad (3)$$

ここで，T：目的達成までの所要時間，

$\gamma : discount factor$

が理想的な状態の評価値となる．つまり，単一目標では，TD型もTS型も基本的に目的達成までの所要時間を状態の評価値としており，TD型では，指数関数で，TS型では，直線で時間の評価を行うという違いになる[7)]．したがって，理想的な状態評価値が学習によって獲得されたとすれば，両者の値は1対1の対応がとれ，状態評価値の最急勾配方向も等しくなる．以上より，以下の議論では，状態評価値のlandscapeを除いて，基本的にTD型の強化学習についても有効である．TS型では，出力の差が時間の経過と比例するため，状態評価値から目的達成までの所要時間が容易に推定できる．したがって，ここではTS型の強化学習を用いた．また，上記の状態評価値の時間変化量，つまり，傾き $(P_{max} - P_{min})/N_{max}$ は，TD型の強化学習では，discount factor γ に相当し，これを適応的に変化させていることになる(文献[7)]参照)．

一方，動作は，動作出力ベクトル $\mathbf{m}$ に試行錯誤のための微小な乱数ベクトル $\mathbf{rnd}$ を加えたものにしたがって行う．そして，それによって変化した状態評価値 p の値から

$$\mathbf{m}_s(t-1) = \mathbf{m}(t-1) + \mathbf{rnd}(t-1)\{p(t) - p(t-1)\} \qquad (4)$$

ここで，$\mathbf{m}_s$：動作信号に対する教師信号

という教師信号を生成してBP法で学習する．これによって，$p(t)$ の値が大きいときの試行錯誤成分 $\mathbf{rnd}(t-1)$ がより強化されることにより，状態評価値がより大きくなる，つまり，より目標に時間的に近づく連続値動作を学習していく．

4.　シミュレーション

4.1　目標物到達タスクと中間層における適応的表現

本論文では，視覚センサ付きの移動ロボットが目標物に到達するというタスクを扱う．そこで，始めに，その基本となるタスクの概要と学習後に目標物の位置が中間層でどのように表現されているかを示す．

Fig. 1にシミュレーションの環境を，Fig. 2にシステムの構成と信号の流れを示す．このロボットは，中心から1.0離れた両脇に車輪を持ち，独立に制御して回転させることによって前後および回転運動ができる．そして，その車輪の上に1つずつ計2つの視覚センサを持っているものとする．各視覚センサは，1次元に配置された多数のセンサセルよりなり，各センサセルはオーバーラップのない放射状の広がりを持つ局所的な受容野を持ち，受容野中で目標物が投射されている面積の割合を0から1の連続値で出力する．センサ全体として180度の視野を有するものとし，視覚センサ上に投射

された物体が占める面積は，その距離に反比例する．ニューラルネットは 3 層で，出力層のニューロン数は 3 個である．そのうち１つが状態評価値として使われ (critic)，残りの２つが車輪の回転速度として使われる (actor)． そして，目標物を Fig. 1中の平面内（$-5 \leq x \leq 5, 0 \leq y \leq 7$）にランダムに置き，ロボットが目標物の中心を通り抜けたときに，目標物を獲得して報酬が得られたものとし，0.4 という教師信号で状態評価値を学習した．逆に，それ以外で目標物が視野から消えた，つまり，目標物が自分よりも後方に行ってしまった場合は罰として，-0.4 の教師信号によって学習を行った．そして，目標物に到達するかまたは見失った場合を１試行とし，再びロボットを元の位置に戻し，目標物の位置を乱数によって決定し，動作と学習を再開した．ただし，学習初期には，目標物の初期位置は，ロボットに近い位置に限り，学習の進行と共に徐々に Fig. 1の範囲まで広げていった．

始めに，すでに文献[7]で示されているが，この簡単なタスクを学習することで，中間層が目標物の位置をどのようにコーディングするかを簡単に示す．ただし，目標物の直径は１とした．まず，Fig. 3のように，各層のニューロン数が 48-20-2-20-3 の５層のニューラルネットを用意した．そのニューラルネットを使って強化学習を行い，３層目の２つの中間層ニューロンが物体の位置をどのようにコーディングしているかを観察した．

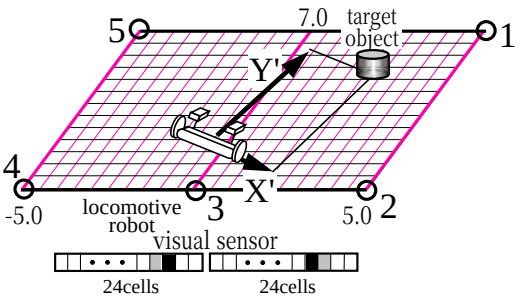


Fig. 1 Simulation environment of “Going to a target” task for a robot with visual sensors

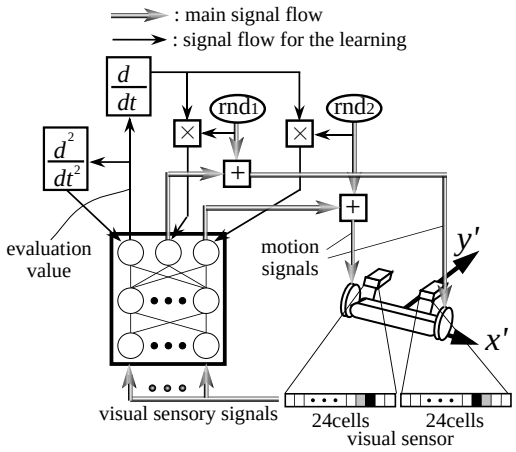


Fig. 2 Reinforcement learning system and signal flow

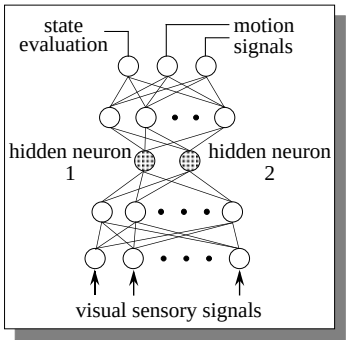


Fig. 3 5 layered neural network to observe the coding of hidden neurons.

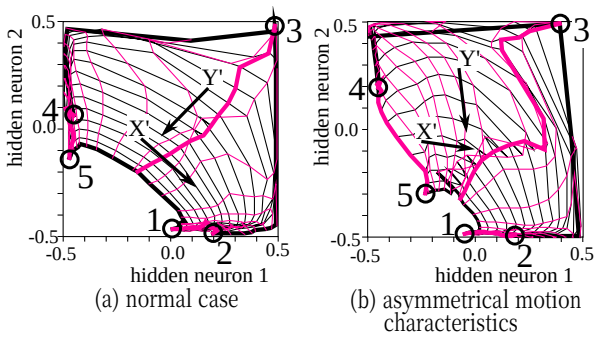


Fig. 4 The coding of the target location in the hidden layer and the change of the coding according to the motion characteristics by reinforcement learning.

Fig. 4に，中間層の２つのニューロンの値を縦軸と横軸にとり，Fig. 1中の３の位置にロボットを固定し，格子上に物体を置いたときの中間層の値をプロットしたものを示す．図中の数字付の○印は，Fig. 1中の○印に物体を置いた場合の中間層の値であり，○３から左下にのびているうすい太い線がロボットの正面を表している．学習前はニューラルネット内の初期重み値を微少乱数としているため，中間層ニューロンの値は，目標物の位置によらず，ともに 0.0 付近であるが，100000 試行後には，Fig. 4(a) のように，中間層ニューロンの値域ほぼ全域を使って物体の位置を表現している．さらに，中間層では，目標物の位置を一様に表現しているのではなく，学習に必要なところを拡大して再構成していることがわかる．また，右側の車輪のみ，回転速度３倍にするという非対称な動作特性を導入すると，Fig. 4(b) のように，中間層の表現は大きく変化した．これより，まわりの環境は全く同じでも，ロボットの動作特性を変化させるだけで空間情報を適応的に表現する能力があることを示している．詳細は文献[7]を参照されたい．

4. 2 可変サイズ目標物とステレオ視

4. 2. 1 タスクの設定と学習の経緯

始めに，前述の目標物到達タスクにおいて，目標物の大きさを試行ごとに可変とし，両眼立体視の機能を，設計者が与えることなく，強化学習だけで獲得できるかどうかを調べた．

目標物の大きさは，直径が 1.0 から 2.0 の範囲で，各試行

の開始時に乱数で決定した．視覚センサセルの数は，左右それぞれ 48 個とし，合計 96 個の信号がニューラルネットに入力される．この場合，目標物に到達するまでの所要時間は，目標物の大きさにはよらず，目標物の位置のみによって決まるが，目標物の大きさが可変のため，目標物の位置を知るためには，両眼立体視を行わなければならない．

学習後に，最大の大きさの目標物を提示したときと，最小の目標物を提示したときのロボットの経路を Fig. 5に示す．また，ロボットから見た目標物の位置に対する評価値の分布を Fig. 6（ロボット中心座標）に示す． これらの図より，ロボットの経路，評価関数共に物体の大きさにあまり依存しないようになっていることがわかる．

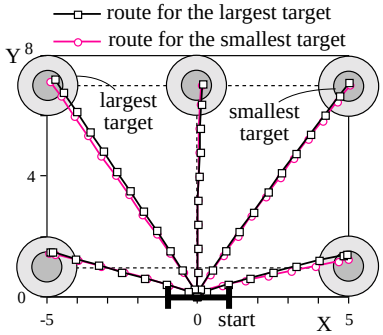
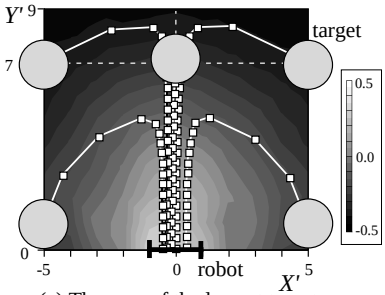
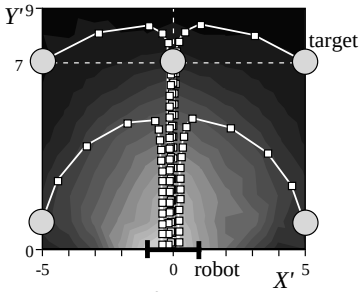


Fig. 5　Sample robot's routes after learning in the cases of the largest target and the smallest target in the robot centered coordinates. Each point is plotted at every 10 time units.



(a) The case of the largest target



(b) The case of the smallest target

Fig. 6　Distribution of evaluation value and robot's routes after learning in the cases of the largest target and the smallest target in the robot centered coordinates. Each point is plotted at every 10 time units.

4.2.2　中間層の表現

つぎに，中間層での物体情報のコーディングを調べた．Fig. 7のように，強化学習を行うニューラルネットに中間層との結合の重み値をすべて０にした出力ニューロンを１個付加し，そのニューロンに対していくつかの学習データを与えて教師あり学習を行う．そして，学習完了後，学習に用いていないテストデータを与えて，強化学習の適用前と後でその出力を比較する．両者とも同じ初期重み値を用いることにより，強化学習の適用による中間層のコーディングの変化，つまり，入力－中間層間の結合の重み値の変化が最終的な出力の差となる．そこで，教師あり学習後の出力を比較することにより，強化学習によって中間層がどのようなコーディングをするようになったのかを推察する．ここでは，教師あり学習時に入力層から中間層への重み値は固定せず，学習によって変化させた．ただし，強化学習後のネットワークに関しては，重み値を固定しても，結果が大きく変化することはなかった．

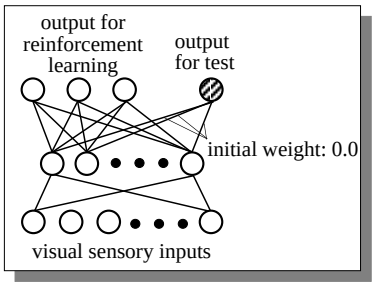


Fig. 7　A neural network to observe the hidden neurons' representation after reinforcement learning.

この節では，中間層における物体の位置のコーディングについて調べた．まず，教師あり学習時には，最大サイズの目標物をランダムな場所に提示し，そのときの視覚信号を入力して，ロボットと目標物までの距離に応じた Fig. 8(a) のような教師信号によってニューラルネットを学習した．学習後，最小サイズの目標物を提示して，提示位置に対する出力の分布を調べた．もし，物体の位置を，視覚センサ上に投射された物体の大きさで判断していれば，小さい物体は遠くにあると判断し，ステレオ視を利用していれば，小さい物体に対しても，正確な位置を出力すると予想される．その結果を Fig. 8(b)(c)(d) に示す．(b) は強化学習適用後の場合，(c) は強化学習時に最大サイズの物体のみで学習を行った場合，(d) は強化学習を行わないで教師あり学習のみを行った場合をそれぞれ示す． この図より，可変サイズで強化学習を行った場合は，出力分布が，与えた教師信号のように滑らかではないものの，その他の場合と比較すると，目標物の位置に対して比較的教師信号に近い値を出力していることがわかる．

教師あり学習時の誤差，および，最大サイズに対して与えた教師信号と最小サイズに対する出力との差を提示位置に対して平均した値が，それぞれ教師あり学習の進行とともにどのように変化したかを Fig. 9と Fig. 10に示す． この図より，

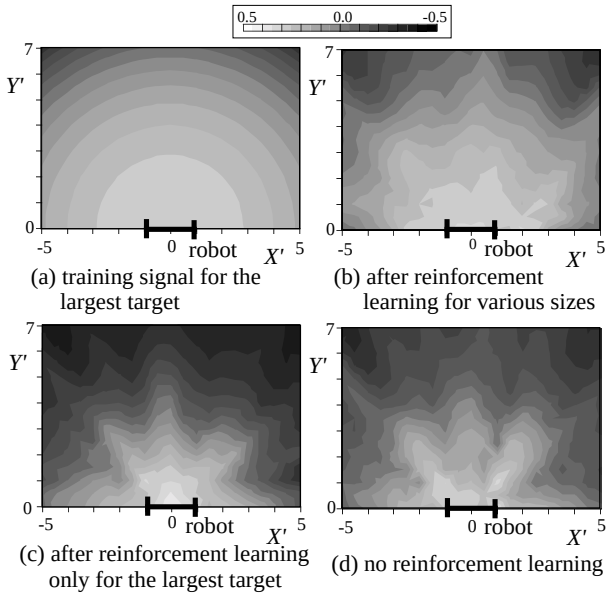


Fig. 8 Output distribution as a function of the target location after the supervised learning for the largest target when the smallest target was presented.

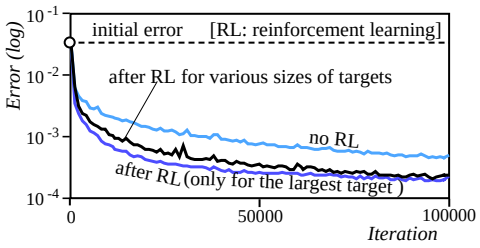


Fig. 9 Learning curve in the supervised learning for the largest object.

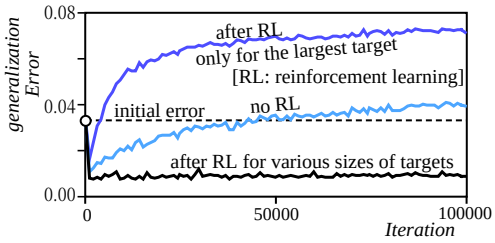


Fig. 10 The change of the difference between the output when the smallest target is presented and the training signal given for the largest target.

教師あり学習の誤差自体は，いずれの場合も，学習の経過とともに減少していることがわかる．一方，最小サイズの目標物を見せたときの出力との差は，強化学習後の場合は減少したが，強化学習を行わなかった場合と，最大サイズの目標物のみで強化学習を行った場合は，いったん下がるものの，その後上昇することがわかった．特に，後者に付いては，教師あり学習を行う前よりも差が大きくなることがわかる．これは，強化学習を通して，主として視覚センサ上の物体の大きさから物体との距離を認識することを獲得したため，強化学習を行わなかった場合と比較して，小さな物体に対し，距離が遠いという認識をより強く行ったものと考えられる．

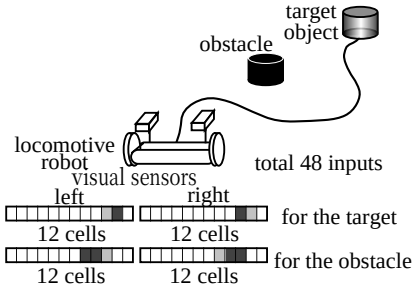


Fig. 11 "Going to a Target" Task with an Obstacle

以上のことから，両眼立体視よりも，視覚センサ上に投射された物体の大小による位置認識を獲得するほうが容易であると考えられる．しかし，可変サイズにして両眼立体視が必要とされる環境では，同じ強化学習をさせるだけで，両眼立体視によって物体の位置認識を行うことを，設計者から与えられることなく獲得しており，状況に応じた認識方法を学習によって獲得する能力があることがわかる．

4.3 障害物回避タスク

4.3.1 タスクの設定と学習の経緯

つぎに，Fig. 11のように，ロボットの前に障害物を置き，目標物への接近と障害物の回避という２つの目的を，ロボットが一つのニューラルネットで学習することができるかどうかを確認した．ここでは，目標物のみを捕えるセンサと障害物のみ捕える視覚センサを仮定した．それぞれ，左右 12 個ずつのセンサセルで構成し，全部で 48 個のセンサ出力を直接ニューラルネットに入力した．また，それぞれのセンサは，他の物体の後ろに隠れても見えることとした． 目標物と障害物の直径は 1.0 とし，２つの物体は，それぞれ各試行ごとに乱数を使って $-5 \leq x \leq 5, 0 \leq y \leq 7$ の範囲で，かつ両者が 2.0 以上離れたところに置いた．そして，ロボットが障害物に当たった場合は，後退しない限り，たとえ動作出力が０でなくても前には進めないこととした．また，衝突に対して罰は与えなかったが，前に進めずにトラップされることによってその状態の評価は下がり，衝突を避ける動作を学習することが期待される．また，目標物を取り込む学習が進むまで，障害物は置かなかった．ニューラルネットは試行錯誤から 4 層とし，中間層ニューロン数は下から，30 個，20 個とし，層間は全結合とした．

167000 試行後 (200000 試行中に，1000 試行ごとにサンプリングした場合の最良解) に，障害物を正面 $(x, y) = (0.0, 3.0)$ に置いた場合の，6ヵ所の目標物に対するそれぞれのロボットの経路を Fig. 12 に示す． 目標物の初期位置によっては，障害物にぶつかって進まなかったり，障害物の前で立ち止まったりする場合があった．図中の点々の領域は，ロボットが立ち止まってしまった場合の目標物の初期位置の範囲，斜線の

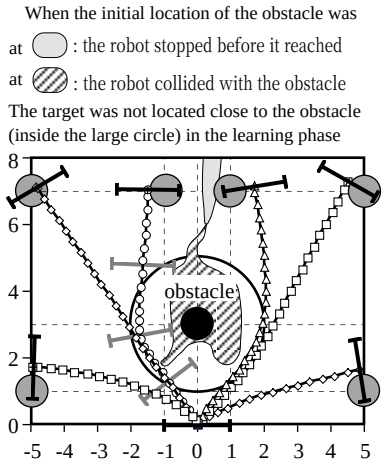


Fig. 12 Comparison of the robot's routes after learning between when an obstacle is located and when no obstacles are located.

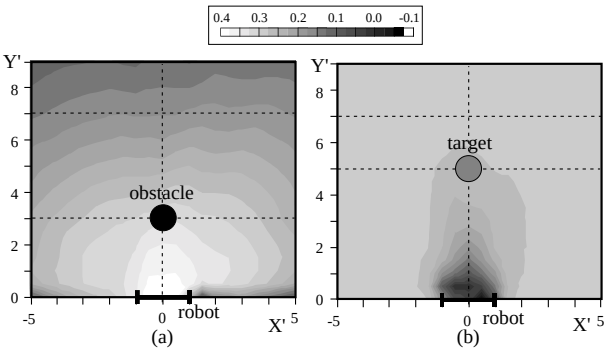


Fig. 13 Distribution of evaluation value as a function of the target location in the robot centered coordinates (a)when the obstacle is fixed at (0.0, 3.0) (b)when the target is fixed at (0.0, 5.0).

領域は，障害物とぶつかって前に進めなかった場合の目標物の初期位置の範囲を示している．これらの領域をのぞいて，ロボットは効率よく障害物を避けて目標物までたどり着くことができている．また，学習過程において，この経路はあまり安定でなく，障害物が正面にあって，目標物が右側に見えても障害物の左側を通って行く場合もあった．これは，ロボットが右左という概念を持ち合わせていないため，障害物の右側を通って目標物に達すると，そのときの学習による汎化から障害物の右側を通りやすくなるためと考えられる．

Fig. 13(a) に，障害物を $(x, y) = (0.0, 3.0)$ に固定した場合の目標物の位置に対する評価関数を示す．この図より，目標物がロボットに近ければ評価が高いこと，さらに障害物の後ろでは評価がわずかながら下がっていることがわかる．Fig. 13(b) に，目標物を $(x, y) = (0.0, 3.0)$ に固定した場合の障害物の位置に対する評価関数を示す．この図より，障害物が目標物の前にいる状態が悪い状態であり，それ以外の場合は，障害物がどこにいてもあまり評価には影響していないことがわかる．これは，障害物は避ければよく，遠くに逃げる必要はないという評価を学習を通して獲得したことを意味しており，合理的であると言える．

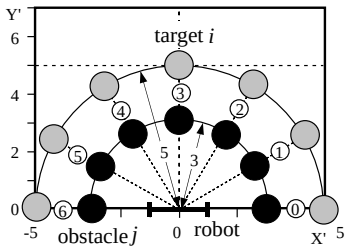


Fig. 14 The target and obstacle locations in the simulation to examine the hidden neurons' representation.

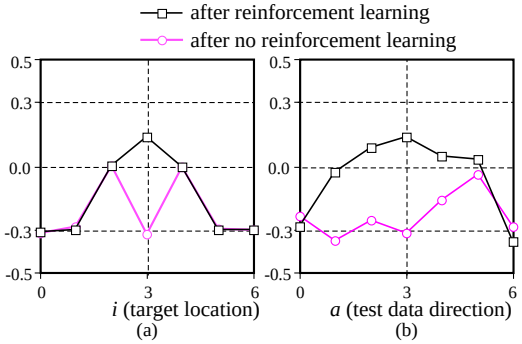


Fig. 15 Comparison of the output after supervised learning between the trained hidden neurons and no trained ones. (a) the output as a function of the target location i when the obstacle is fixed at $j = 3$. (b)the output as a function of the location a of the test data.

4.3.2　中間層の表現

ここで，前節と同様に，強化学習を行った後に教師あり学習を行って，中間層の表現を観察した．このタスクでは，目標物が障害物の後ろにあるという状態のときに通常の目標物到達とは違った動作が要求される．障害物の後ろに目標物があるという状態は，両者がロボットの正面にある場合，右のほうにある場合，左のほうにある場合などいろいろなケースがある．それぞれの場合で，入力信号は全く違ってくるが，状態評価値が低くなることや避けるという動作をすることなどの知識はある程度共通に使うことができるはずである．そこで，ニューラルネットが，強化学習を通してそのような状況を個々の状態としてではなく，同じ状況として認識することができているかどうかを調べる．これをここでは，“知識の抽象化” と呼ぶ．

まず，Fig. 14 に示したように，目標物の位置をロボットから 5 離れた場所に 7ヵ所，障害物の位置をロボットから 3 離れたところに 7ヵ所用意し，その中からそれぞれ乱数によって目標物の位置 i，障害物の位置 j を選択する．そして，そのときの視覚イメージをニューラルネットに入力する． 教師信号は，障害物と目標物が同一方向の場合に 0.3，方向が 1 つずれている場合には 0.0，それ以外の場合には -0.3 として学

習を行った．ただし，$i=j=a$ (ただし，aは0から6のいずれか) の場合は学習させず，学習後にテストデータとして用いた．学習で用いたパターンの中で，$i=a$ または $j=a$ の場合，教師信号はすべて0.0以下であるため，入力信号に対する汎化を考えれば，$i=j=a$ の場合も，出力は0.0以下になることが予測される．この結果を強化学習後と強化学習を適用しなかった場合で比較した．ニューラルネットの結合の重み値の初期値は同じにした．Fig. 15 (a) に，$a=3$ の際に，障害物の位置を3に固定した場合の目標物の位置に対するニューラルネットの出力を示した．これより，強化学習を行わなかった場合には，$i=j=3$ の出力は0より小さくなっており，強化学習を行った後では，0より大きくなっていることがわかる．強化学習を行わなかった場合は，前述の入力信号に対する汎化から予想した結果と一致したものとなっている．しかし，強化学習後の結果は，入力信号空間上での汎化からは説明できない．これは，強化学習を行うことにより，中間層において，目標物が障害物の後ろに隠れる状態を他と区別して表現していた結果であると考えられる．これは，入力が違っても教師信号が近ければ中間層の表現が似てくるというニューラルネットの性質によるもの (文献[15]参照) と考えらる．これは，知識の抽象化としてとらえることができ，中間層を共有することで別のタスクにおいてもその知識を容易に利用できる可能性を示唆している．

つぎに，テストデータの位置aを変化させたときの出力のようすをFig. 15 (b) に示す．この図より，テストデータが視野の端にあるときは，強化学習を行った場合と行わなかった場合でその出力に大きな差異はなく，中間層における目標物が障害物の後ろに隠れるという識別に関する汎化は，視野の端のほうでは有効でないことがわかる．また，強化学習を行っていない場合の出力がかなり変動しているが，これは，ニューラルネットの結合の重み値の初期値によるものである．

5. 視覚センサ入力の効果

5.1 視覚センサ信号による状態の爆発

第1章で述べたように，視覚センサ信号は信号数が多く，状態の爆発を起こして学習が遅くなるのではないかと心配される．しかし，各センサ信号はランダムには変化せず，隣同士のセンサセルの出力は近い値であることが多く，冗長性が高い．したがって，センサセルの数をnとしても，状態数は，単純にn乗のオーダーで増えることにはならない．また，多数の入力に対し，並列演算ができるようにハードウエア化すれば，計算時間はさらに減少する．また，前節で示したように，いったん学習して中間層の表現を獲得すれば，つぎに別のタスクを学習する際には，中間層の表現の上で学習することが可能になる．以上より，信号数増大によって学習が遅くなるという心配はあまり致命的ではないと考えられる．

5.2 強化学習＋ニューラルネットの学習不安定性の指摘

Boyanらは，ニューラルネットと強化学習の組み合わせが学習の不安定につながるとの結果を示している[8]．これに対し，GordonやSuttonは，CMACなどの入力信号を局所化する手法を用いて，on-lineで学習させることによって，学習の不安定を回避でき[12)13)]，逆に，シグモイド型のニューラルネットなどでは学習が発散する場合があることを示している[12)]．ただし，常に発散するわけではない．シグモイド関数は非線形性が弱い関数であり，それを出力関数とするニューラルネットは，ステップ関数などの強い非線形関数近似を苦手とする．したがって，強い非線形性が要求されるタスクでは，Gordonの指摘のように学習が不安定になると推測される．実際，Boyanらの示した車が山を登るタスクも強い非線形性が要求される．

本論文では，ロボットの学習ということで，on-line学習を前提としているが，さらに，視覚センサ信号を直接シグモイド型のニューラルネットに入力している．視覚センサをはじめとして，多くのセンサは，多数の局所的な受容野を持ったセンサセルによって構成されている．たとえば，視覚センサ上にある物体の位置は，大域的な連続値で表現することができるが，視覚センサは，それを局所化する働きをしているととらえることができる．したがって，視覚センサ信号は，すでに局所化されているため，シグモイド関数を用いたニューラルネットでも，局所化された信号を自由に組み合わせることにより，容易に強い非線形性を実現することができる．

ところが，NGnet(Normalized Gaussian Network) を含むRBF (Radial Basis Function) やCMACベースの学習[14)13)]では，局所化した情報からそのまま出力を求めているため，その内部に大域的な情報を表現していない．したがって，テーブルを構成している状態に近い．一方，視覚センサ信号をニューラルネットに入力すると，ニューラルネットの中間層で大域的な表現を獲得することができる[15)]．したがって，空間認識など様々なタスクで利用できる知識は，いったん学習されると，中間層を共有することで，つぎのタスクでは，その中間層で表現されている大域的な空間上で汎化がきく．これによって，0から学習する必要がなくなり，飛躍的に学習時間が短縮することが期待される．

上記の指摘以外にも，Bairdも学習の発散を指摘しているが[16)]，これは，off-lineを前提とした話となっている．さらに，Tsitsiklisらの指摘[17)]では，非線形関数近似を用いると発散することを，例を挙げて示しているが，これは，指数関数のような非線形関数を用いた場合の指摘であり，シグモイド関数を用いたニューラルネットへの指摘にはあてはまらないと考えられる．また，ニューラルネットを用いると，学習が収束する保証がないとの指摘もあるが，収束しないという指摘ではない．以上の議論から，筆者らは，むしろ，ニューラルネットを積極的に使用していくべきであると考える．

5.3 視覚センサ信号入力 vs 大域的信号入力

本節では，視覚センサ信号の局所性が学習の高速化，安定化に有効であることを確認するため，視覚センサ信号を直接入力した場合と，ロボットから見た目標物の前後と左右の2つの相対位置を入力とした場合の学習について比較した．タ

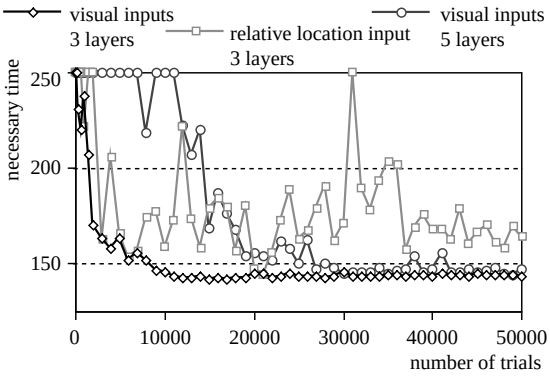


Fig. 16　Comparison of the learning with respect to the form of the input

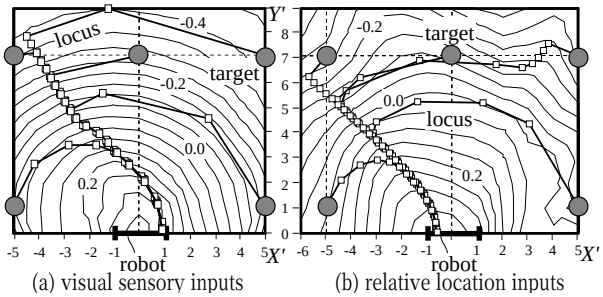


Fig. 17　Comparison of the state evaluation function as a function of the relative target object location and the locus of the robot on the robot-centered coordinates after 50000 trials with respect to the input form.

スクは，4.1 節で述べた非対称動作特性の場合と同じである．このタスクも，目標物を捕らえられるか捕らえられずに通りすぎてしまうかの境界では，状態評価値および動作が不連続に変化することが要求され，ロボットからみた目標物の位置という大域的な情報からは強い非線形性が要求される．ここでは，(1) 視覚信号入力&3 層ネットワーク，(2) 視覚信号入力&5 層砂時計型ネットワーク，(3) 前後，左右の２つの相対位置入力&3 層ネットワークの３つの場合について比較した．3 層ネットワークの中間層ニューロン数は 20 個とし，5 層砂時計型ネットワークの中間層ニューロン数は，4.1 節の場合と同様に 20-2-20 とし，真ん中の中間層より上の部分では，(3) の相対位置入力の場合と全く同じ構造とした．

Fig. 16に，各試行回数学習後のニューラルネットを用いて，5ヶ所に目標物を置いた場合の所要時間の平均値をプロットした．ただし，それぞれ 250 単位時間で打ち切っているため，全く到達できないような状態でも，所要時間が 250 単位時間となっている． これより，視覚センサ信号を直接入力し，3 層ネットワークに入力した場合が学習が最も早く，かつ学習が進んだときの所要時間も小さく，安定である．一方，相対位置を入力した場合は，学習が非常に不安定である．また，同じ視覚センサ信号を入力した場合でも，砂時計型では，層数が多く，真ん中の中間層ニューロン数が２つしかないため，学習が極端に遅いが，相対位置を入力した場合と比較して安定している．砂時計型では，空間情報を，真ん中の中間層で，視覚センサ信号から２つの連続値に変換しなければならないが，Fig. 4(b) のように，不連続性が要求されるところは拡大して表現するなど，その表現を学習によって適応的に獲得できる分，表現が固定されている相対位置入力の場合より学習が安定しているものと考えられる．以上より，視覚センサ信号から前処理で目標物の相対位置を計算して入力することは，強化学習の能力を阻害しており，逆に，学習を不安定化させていることがわかる．

また，学習によるロボットの動作の変化を観察すると，視覚センサ信号を入力し，3 層ネットに入力した場合は，Fig. 17 (a) のように，まず回転して目標物が自分の左前方に見えるようになったら前進するという行動をとり，回転から前進に切り替わる位置や目標物を取り込む位置は試行数によってあまり変化しない． しかし，相対位置入力の場合は，Fig. 17(b) のようになるなど，回転から前進に切り替わる位置や目標物を取り込む位置が学習によって大きく変化し，安定しなかった．この傾向は，学習定数を小さくしても見られた．

5.4　他のセンサ信号への適用

本論文では，典型的な例として視覚センサの場合を挙げているが，上記の議論は，視覚センサ信号に限らず，皮膚の触覚センサなどをはじめとする局所的な受容野を持つセンサセルが多数集まって構成されているセンサ一般に関して適用可能である．また，聴覚においても，Jeffress の音源定位における局所化のモデル [18] が生理学的にも確認されてきていること [19] や，蝸牛の基底膜がバンドパスフィルタの役割をし，周波数空間での情報の局所化が見られるという点 [20] はたいへん興味深いところである．

6.　学習時間に関する考察

本論文では，視覚センサ信号入力による学習の高速性を指摘しているが，実際には，本論文中のシミュレーションでは，非常に単純な視覚センサを仮定しているにもかかわらず，なお数万から数十万回の試行を必要としている．試行錯誤方法や，学習方法自体にまだまだ改良点があると考えられる．ただし，本論文では，全く知識のない状態から学習を始めているため，ある程度学習に時間がかかるのは仕方がない面もあると考える．また，中間層に蓄えられた空間認識は様々なタスクで必要とされるため，いったん獲得すれば別のタスクで利用できることも考慮すべきである．しかし，現実的には，ロボットに適用する際には，0 から学習することはあまり得策ではなく，馬が生後すぐに歩けるように，何らかの知識をあらかじめ付与し，学習を高速化させることも必要であると考える．ただし，その知識によって適応性，柔軟性が阻害されないようにするため，筆者らは，ニューラルネットの初期値として知識を付与することを提案する．これによって，学習が加速される上，必要に応じてあらかじめ与えられた知識をロボット自身で容易に変更することができる．

7. おわりに

ロボットの自律学習のために，センサからモータまでを階層型ニューラルネットで構成し，強化学習を行う，Direct-Vision-Based 強化学習を提案した．ニューラルネットを用いることで，全体をシームレスな構造とし，合目的的，調和的かつ効率的に学習させることができることを主張した．

そして，両眼立体視を必要とする可変サイズ目標物到達タスク，障害物回避タスクなどに適用し，この学習が比較的難しいタスクにも対応する能力を持っていること，さらに，中間層において，あらかじめ知識を与えることなく，空間認識に関する知識の抽象化が行われることを示した．

最後に，個々の視覚センサセルが大域的な空間情報を局所化する働きを持っていることを指摘し，それが，学習を高速かつ安定にすることをシミュレーションによって示した．

謝辞

本研究の一部は，文部省科学研究費重点領域研究「創発システム」(No. 264)，基盤研究 (No. 09750484) および学術振興会未来開拓学術研究推進プロジェクト「生物的適応システム」(JSPS-RFTF96100105) の補助の下で行われました．また，査読者の方から多くの有用なコメントをいただきました．ここに謝意を表します．

参考文献

1) G.J. Tesauro : Practical Issues in temporal difference learning, *Machine Learning*, **8**, pp.257–277 (1992)
2) R.H. Crites & A.G. Barto : Improving elevator performance using reinforcement learning, *Advances in Neural Info. Processing Sys.*, MIT Press,**8**, pp.1017–1023 (1996)
3) 浅田稔 他 : 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, **13**, 1, pp.68–74 (1995)
4) M. Asada et al: Purposive Behavior Acquisition for a Real Robot by Vision-Based Reinforcement Learning, *Machine Learning*, **23**, pp.279–303 (1996)
5) 浅田稔 他 : 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, **15**, 6, pp.886–892 (1997)
6) 西野哲生, 柴田克成, 岡部洋一: 遅延強化信号による視点移動の学習, 信学技報, NC96-135, pp.171–178 (1997)
7) 柴田克成, 岡部洋一 : 時間軸スムージング学習, 電気学会論文誌 C 分冊, **117-C**, 9, pp.1291–1299 (1997)
8) J.A. Boyan & A.W. Moore: Generalization in Reinforcement Learning: Safely Approximating the Value Function, *Advances in Neural Information Processing Systems*, MIT Press, **7**, pp. 369-376 (1995)
9) D. E. Rumelhart, G. E. Hinton & R. J. Williams: Learning Internal Representations by Error Propagation, *Parallel Distributed Processing*, The MIT Press, **1**, pp. 318-362 (1987)
10) A. G. Barto, R. S. Sutton & C. W. Anderson : Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems, *IEEE Trans. SMC*, **13**, pp. 835-846 (1983)
11) 柴田克成, 岡部洋一 : 時間軸スムージング学習を用いた局所センサ信号の統合と空間情報の抽出, 日本神経回路学会誌, **3**, 3, pp. 98-105 (1996)
12) G. J. Gordon: Stable Fucntion Approximation in Dynamic Programming, *Proc. of the 12th Int'l Conf. on Machine Learning*, pp. 261–268 (1995)
13) R. S. Sutton: Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding, *Advances in Neural Information Processing Systems*, MIT Press, **8**, pp. 1038–1044 (1996)
14) 森本淳, 銅谷賢治: 強化学習を用いた高次元連続状態空間における系列運動学習 -起き上がり運動の獲得-, 電子情報通信学会論文誌, *J82-D-II*, No. 11, pp. 2118-2131 (1999)
15) K. Shibata and K. Ito: Reconstruction of Visual Sensory Space on the Hidden Layer in Layered Neural Networks, *Proc. of Int'l Conf. on Neural Information Processing (ICONIP) '98*, **1**, pp. 405-408 (1998)
16) L. Baird: Residual Algorithms: Reinforcement Learning with Function Approximation, *Proc. of the 12th Int'l Conf. on Machine Learning*, pp. 30-37 (1995)
17) J.N. Tsitsiklis & B.V. Roy: An Analysis of Temporal-Difference Learning with Function Approximation, *IEEE Trans. Automatic Control*, **42**, No. 5, pp. 674-690 (1997)
18) L. A. Jeffress: A Place Theory of Sound Localization, *J. Comp. Phsiology and Psychology*, **42**, pp. 35–39 (1948)
19) 力丸裕: 聴覚認知, 脳科学大事典, 朝倉書店, pp. 147–157 (2000)
20) D. D. Greenwood: A Cochlear Frequency-Position Function for Several Species – 29 years later, *J. of Acoustic Society of America*, **87**, No. 6, pp. 2592-2605 (1990)

[著 者 紹 介]

柴 田 克 成（正会員）

1989 年 東大大学院工学系研究科機械工学専攻修士課程修了．1989 年 (株) 日立製作所 に入社．1992 年 10 月 同社退職．1993 年 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 年 東大先端科学技術研究センター助手．1997 年 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 年 大分大学工学部電気電子工学科講師．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））．

岡 部 洋 一

1972 年 東大大学院工学系研究科電子工学専攻博士課程修了．1972 年 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（工学博士）．

伊 藤 宏 司（正会員）

1969 年 名大大学院工学系研究科応用物理工学専攻修士課程修了．1970 年 同大工学部自動制御研究施設助手．1979 年 広島大学工学部電気系助教授．1992 年 豊橋科学技術大学情報工学系教授．1996 年 東工大大学院総合理工学研究科教授．主として，運動制御，ロボティクス，マンマシンインターフェースの研究に従事（工学博士）．