視覚センサ付き実ロボットによる箱押し行動の獲得
－強化学習によるセンサーモータ間トータル機能獲得への第一歩－

Acquisition of Box-Pushing Behavior in a Real Mobile Robot with a Visual Sensor
- The First Step of the Acquisition of the Total Function from Sensors to Motors by Reinforcement Learning -

○柴田克成（大分大学）　飯田大（大分大学）

Katsunari Shibata, Dept. of Electrical and Electronic Engineering, Oita University, 700 Dannoharu, Oita 870-1192, Japan

Masaru Iida, Oita University

In this paper, it was shown for the first time in the world that a real robot with a camera could learn appropriate actions only by the approach of reinforcement learning for the total function. Aiming at the acquisition of not only the motion planning, but also the total function from sensors to motors, raw visual sensor signals are the inputs of a layered neural network; the neural network is trained by Back Propagation using the training signal that is generated based on reinforcement learning. In other words, no image processing, no control methods, and no task information are given at premise. The box-pushing task employed in this paper is rather difficult for the reason that (1)as many as 1536 monochrome visual signals and 4 infrared signals are the inputs, (2)not only the center of gravity of the box image, but also the direction, weight and sliding character of the box should be considered, and furthermore, (3)a reward is given only when the robot is pushing the box. It was also observed that the neural network obtained global representation of the box location in the hidden layer through the learning.

Keywords: reinforcement learning for the total function, box pushing, neural network, mobile robot, global representation

１. はじめに

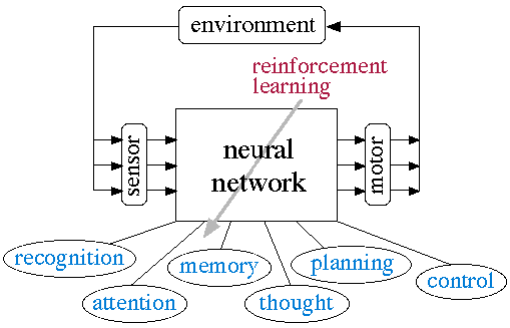
環境からたくさんの情報を得るために、たくさんのセンサ信号を出力する視覚センサを利用するロボットが多くなっている。しかし、学習に重点を置いて開発されたロボットでさえ、視覚センサ信号からタスク達成のために有効な情報を抽出するための画像処理・認識のプログラムについては、あらかじめ与えられることが当然のこととされてきた。

たとえば、浅田らのサッカーロボットがシュート行動を学習する研究[1]では、ロボットが捉えた画像からボールの位置や大きさ、ゴールの位置や大きさ、向きなどを抽出している。これらは、タスク達成に非常に重要な情報であるが、ロボットがどうやってそのような情報を重要であると認識し、そのような情報を抽出するようになるのかというところまでは踏み込んではおらず、そのような情報を抽出する機能をあらかじめロボットに組み込んでいる。

これらの背景には、(1)多数のセンサ信号の処理を学習によって獲得させることは効率が悪い、(2)必要とされる基本的な画像処理はタスク間であまり変わらないので、学習の必要性が小さい、(3)センサからモータまでの過程を機能モジュール分割し、個々のモジュールを高機能化するという伝統的な考えから、画像処理・認識は行動計画とは全く別の処理であると考えられていることなどがベースにある。このような機能モジュール分割の傾向は、脳研究においても見られるが、全体を一度に考えようとしても難しく、どう手をつけて良いかわからないということが理由ではないかと想像される。

ところが、機能モジュールに分割してしまうと、各モジュールの処理の変化が全体としてどういう影響を与えるかを知り、目的にかなった機能を実現することが難しく、また、各機能モジュール間の調和をとることも難しい。われわれ生物の学習から考えられた強化学習は、知覚－行動ループに基づいて行われる自律的な学習であるが、ループの存在が前提であり、個々のモジュールをループから切り離して学習させることは難しい。また、脳研究での知見を見ても、領野の境界部では、中間的な働きをするニューロンが見られるなど、脳の中でも機能的な境界が必ずしもはっきりしているわけではない。これらより、著者らは、われわれの脳の処理は、機能モジュールごとに完全に分割できるものではなく、センサ信号はモータ信号を生成するまで徐々にその表現が変化するものであり、その間に発現するさまざまなレベルの抽象表現がわれわれの知能の原点なのではないかと考えている。そして、ロボットにもそのような機能を搭載することが、柔軟で高い知能を有するロボットの実現につながると考えている。

トータル機能学習のための強化学習は、従来の知能ロボットのように、人間が持っている知識をできるだけロボットに与えるというアプローチとは全く逆で、ロボットのようなシステムに強化学習を適用する際に、あらかじめ与える知識を



Fjg. 1　Reinforcement learning for the total function.

できるだけ少なくし、できるだけ学習によって機能を獲得させるアプローチである[2][3]。具体的には、図１のように、階層型ニューラルネットによってセンサからモータまでを分割することなく構成し、そこに、視覚センサ信号を、画像処理を行うことなく入力し、出力をそのままモータ信号として用いる。これを強化学習に基づいて学習することにより、センサ信号を入力として理想的なモータ信号を出力とする関数が学習され、その結果として、ニューラルネットの中に、単なる行動計画だけでなく、モータ信号生成のために必要な、センサからモータまでの、認識、記憶から制御までも含むトータルな機能が、学習を通して獲得されると期待される。そして、このときニューラルネットの内部に現れる抽象的な情報こそが、従来手が付けられなかったわれわれの高次機能を説明する手がかりになるのではないかと期待している。

筆者らは、シミュレーションによって、視覚センサを持つロボットが目標物へ到達するタスクを学習できることを示した[3]。そして、ニューラルネット内部に、局所的な情報を表現している個々の視覚センサ信号を統合して、物体の位置などの大域的な空間情報を表現していることを確認した。また、その表現は、ロボットの動作特性などによって適応的、合目的的に変化することも確認された。さらに、障害物を置いたタスクでは、目標物が障害物の後ろに隠れるという状態を区別して表現していることがわかった。このように、単に強化学習で学習させるだけで、ある程度抽象的な情報を内部に獲得できるのである。

その後、小型の CCD カメラを搭載した Khepera という小型の実移動ロボットを使って、本当にそのようなことができるのかを実験してきた[4]。そして、直方体の筒状の箱を押したときだけ報酬を与え、箱の位置だけではなく、向きや床に対する滑り方も考慮した近づき方、押し方を学習するというある程度難しいと考えられるタスクを、このアプローチで、シミュレーションによる学習なく、実機のみで学習させることに世界で初めて成功した[5]ので報告する。

２．トータル機能学習のための強化学習

前述のように、センサからモータまでを一つの階層型ニューラルネットで構成し、センサ信号を直接入力する。そして、出力の一つを critic(状態評価)、残りの出力を actor(動作信号)として用いる。学習は TD(Temporal Difference)型の学習を用い、まず、TD 誤差 $\hat{r}_t$ を

$$\hat{r}_t = r_t + \gamma P(\mathbf{s}_t) - P(\mathbf{s}_{t-1}) \qquad (1)$$

ただし、r_t：報酬、γ：割引率、$P(\mathbf{s}_t)$：critic の出力、$\mathbf{s}_t$：時刻 t の状態、と計算し、１時刻前の critic の出力に対し、

$$P_{s,\,t-1} = P(\mathbf{s}_{t-1}) + \hat{r}_t = r_t + \gamma P(\mathbf{s}_t) \qquad (2)$$

を教師信号として学習する。一方、動作信号は、actor の出力 $\mathbf{a}(\mathbf{s}_{t-1})$ に試行錯誤のための乱数成分 $\mathbf{rnd}_{t-1}$ を加えて実際の動作信号とする。そして、

$$\mathbf{a}_{s,\,t-1} = \mathbf{a}(\mathbf{s}_{t-1}) + \hat{r}_t \cdot \mathbf{rnd}_{t-1} \qquad (3)$$

を教師信号として actor を学習する。ニューラルネットは(2)(3)の教師信号により、誤差逆伝搬法（BP 法）で教師あり学習を行う。これにより、critic は、報酬がない場合は、報酬が得られるまで時間とともに指数関数的に上昇するように学習され、actor は、critic の値がより上昇するように学習が進む。そして、ニューラルネットの中間層に抽象表現が獲得されることを期待する。また、より抽象化した表現の獲得には層数を増やすこと、記憶等を扱うためにはリカレント構造とすることなどが必要である[2]。

３．実験

3.1 実験の設定

実験環境を図２に示す。ロボットの行動領域は白い紙を敷いた 70cmx70cm の領域で、まわりも高さ 10cm の白い紙の壁で覆い、上から蛍光灯で照らした。実験に使われるロボットは、図３に示すような、小型の移動ロボット (AAI 社, Khepera)で、広角レンズで 114 度に広げた視野を持つ CCD カメラ(KEYENCE, CK-200)を搭載した。カメラの特性で、画像の中央部は明るく、周辺部が暗く、また、画像の右端や左端では、画像が歪んでおり、元々直線のものでもカーブして見える。カメラによって得られた画像は、PC 上のキャプチャーカードでキャプチャーする。画像は、カラーの 320x240 画素であるが、計算機の能力のため、画像をグレースケールに変換し、画像の下半分を 5x5 の領域ごとに平均をとった値、全部で 64x24=1536 個の信号を、ニューラルネットが学習しやすいように、それぞれ、最も明るいときを 0.0、最も暗いときを 1.0 と明暗反転させて正規化し、ニューラルネットへ入力した。また、ロボットが前面に持つ４つの赤外線センサからの信号も 0.0 から 1.0 に正規化してニューラルネットへ

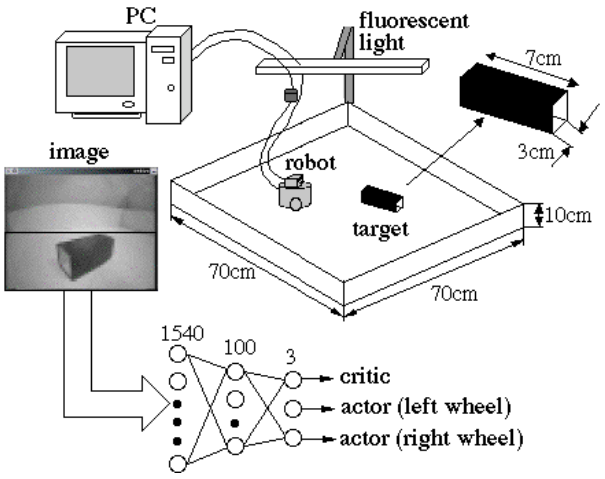


Fig. 2　Experimental Environment

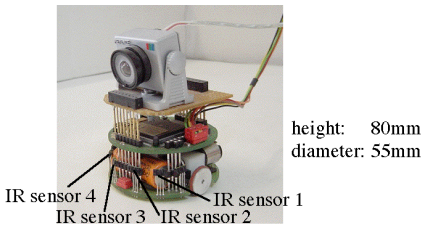


Fig. 3　A small mobile robot named Khepera with a CCD camera.

入力した。センサのちょうど正面に箱がある場合に 1.0 となる。箱は、30mmx70mmx30mm の紙で作られた細長い直方体の筒状で、外は黒、中は白となっており、30mmx30mm の小さい側面からは内部の白い部分が見えるようになっている。

ニューラルネットは３層とし、ニューロン数は、入力層 1540、中間層 100、出力層 3 とした。中間層－出力層間の初期重み値はすべて 0.0 とし、入力層－中間層間は-0.1 から 0.1 の範囲の一様乱数とした。中間層、出力層のニューロンの出力関数は-0.5 から 0.5 の値域のシグモイド関数とし、critic として用いる出力には、0.5 を足したものを実際の値として用いた。つまり、式(1)(2)における P は、実際の出力に 0.5 を足したものとし、逆に P_s から 0.5 を引いたものを実際の教師信号とした。報酬に関しては、二つの赤外線センサ(図３の No.2 と No.3)が、とりうる値の最大値となったとき、つまり箱がロボットのほぼ正面ですぐ目の前に来たときで、かつ両車輪の動作信号が正のときだけ、0.018 という小さい報酬を与えた。また、見失った時には critic の出力に対し、P_s を 0.1、つまり、-0.4 を教師信号として学習した。ロボットが 10 ステップの間報酬を獲得し続けるか、箱を見失うか、スタートから 50 ステップが経過したときに１試行を終了させた。

３つの出力のうち、２つが actor であり、それぞれ、右車輪と左車輪の動作信号の生成に用いた。２つの出力にそれぞれ加えた試行錯誤のための乱数成分は、一様乱数を３乗し、値域を-0.1 から 0.1 としたものを用いた。Khepera の動作信号は整数値で与える必要があるため、actor の出力にこの試行錯誤成分を加えた後、８倍し、それを四捨五入して-3 から 3 の範囲の整数値とした。この値は、RS232C を介してロボットに送られる。学習初期は、0.05 より小さい乱数は、四捨五入すると動作信号が 0 となってしまうため、用いなかった。教師信号が-0.4 より小さいとき、0.4 より大きいときはそれぞれ-0.4、0.4 とした。

次に、各試行開始時のロボットの初期位置について説明する。学習の第１ステージでは、画像上に設定された下辺が上辺より短い小さな台形のエリアからランダムに１点を選び、画像を二値化して得られる箱の(黒い)部分の重心がそこに来るように、あらかじめ与えたプログラムによってロボットを動かした。そして、箱の重心と選んだ点との差が上下、左右ともに１ピクセル以内になったときから学習を開始した。学習の初期には、箱の長辺のちょうど前にロボットが来るように、台形のエリアを画像の下の方でかつ非常に小さく設定し、学習の進行とともに、徐々にその台形エリアを広げていった。ほとんどの場合、ロボットは箱の長辺に向かっており、ロボットの進行方向と長辺が 90 度からあまり離れなかった。そして、学習の第２ステージでは、まず、第１ステージと同じ方法でロボットの位置を決定し、半分の試行においては、その後さらに、ロボットを動かして、進行方向と箱の長辺との角度も変化させた。変化させる角度は、やはり学習の進行とともに大きくしていった。箱が壁にぶつかりそうな時には、試行終了直後に実験者が手で箱を動かした。

すべての学習は、実際のロボットを用いてオンラインで行い、シミュレーションによる学習は一切行わなかった。図４に、ロボットが学習する時のタイミングチャートを示す。１ステップは 320msec とした。各プロセスの実行時間は、おおよそ図４の通りである。”capture”には、モノクロ画像への変換、および、5x5 のピクセルの階調値の平均を求める時間も含まれる。”learning” には、学習のために時間をさかのぼったときの出力の再計算のための計算も含まれる。図４からわかる

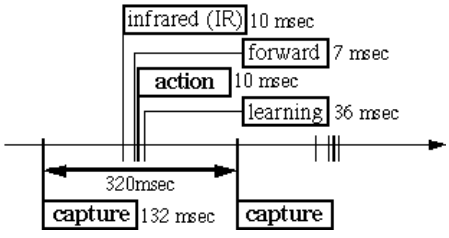


Fig. 4　Timing chart of each event

ように、動作信号出力は、前後の画像のキャプチャの時間の真ん中あたりである。これは、言い換えると、TD 誤差は、実際には、現在の動作信号と一つ前のステップでの動作信号の両方が影響することになる。

3.2 結果

ロボットは実験開始後すぐに前進することを学習した。そして、300 試行のあたりから、箱の位置によって回転する行動が観察された。図５は、5000 試行後のロボットの行動の２つの例を示す。画像処理、制御、タスクに関する知識を一切ロボットに与えていないにもかかわらず、ロボットは箱に到達し、押し続けていることがわかる。ロボットが前進することを学習したのが、予想に反して非常に早い時期であったが、これは、二つの車輪への動作信号がともに正のときだけ報酬を与えたことが理由として考えられる。

ロボットの動作は、箱の位置だけでなく、箱の向きによっても変化した。図６に示したように、箱の位置は同じで向きを変えて、ロボットの行動を観察した。その結果の、ロボットの軌跡とロボットが捉えた画像の変化、および、二値化した画像の黒い部分、つまり箱の重心の変化を図７に示す。

箱が図 6 (a)のように置かれたときは、図 7 (a)のように、ロボットは最初まっすぐ進み、その後反時計方向に回転し、箱の長辺の真ん中あたりに到着した。一方、箱が図 6 (b)のように置かれたときは、ロボットは、図 7 (b)のように、最初少しだけ反時計方向に回転し、それからまっすぐに進んだ。そして、後半わずかに時計方向に回転し、箱の右端に到達した。図５(b)が、このときに上から見た写真である。

この行動の違いは以下のように考えられる。(b)の場合は、もしロボットが箱の正面から長辺の真ん中に向かって近づくと、大回りになり、到着するまでに時間がかかる。また、もしまっすぐ進んでいれば、箱の端が赤外線センサに引っかかり、ロボットが箱に対して相対的に回転することができない。さらに、もし、近づく角度が小さければ、赤外線センサはすぐには最大値にならない。したがって、ロボットは、最初に反時計回りに回転し、それからある程度、長辺に対し角度を保ちながら近づいて行った。ロボットが箱に触った後は、ロボットが箱の端を押すことによって箱が回転し、ロボットの正面に、箱の長辺の真ん中が来るようになっている。このように、ロボットは、学習によって視覚センサ信号から箱の向きも抽出するようになり、さらに、経験を通して、箱の滑りも考慮した行動を獲得したと言える。

次に、学習後の箱の位置に対する critic と actor の出力分布を図８に示す。このとき、箱の位置を 7x8=56 カ所変化させて画像を取り込み、入力として用いた。ただし、箱の長辺とロボットの進行方向は常に直交するように置いた。critic の方は、箱の位置が近いほど値が大きくなっていることがわかる。一方、actor の方は、物体が右に見えると左の車輪、左に見えると右の車輪の出力が大きくなっていることがわかる。

図９に、各入力層ニューロンから２つの中間層ニューロンへの結合の重み値を示す。最初の中間層ニューロン(No.32)は、critic の出力に対して最も大きな重み値(絶対値)で結合しており、２つめのニューロン(No.34)は、actor に対して、最も大きな重み値で結合している中間層ニューロンである。一見、入力層(画像信号)との重み値の分布はランダムに見えるが、よく見ると、No.32 の中間層ニューロンは、画像上部の信号に対しては重み値が正のところが多く、ちょうど目の前に箱があるときに投射されるあたりで、負の重み値が多いことがわかる。一方、No.34 の中間層ニューロンは、画像の右側に該当する入力に対する重み値が大きくなっていることがわかる。この両ニューロンへの信号の重み値が一見ランダムのように見えるのは、図にはないが、冗長性のため、初期重み値における値の大小もある程度保持しているからであった。これは[6]と同じ結果である。表１に、学習前後の両中間層ニューロンへの重み値と入力信号の画像上の位置 i,j との相関を示す。No.32 のニューロンは j との相関が強く、一方、No.34 のニューロンは i との相関が強くなっていることがわかる。

図１０に、学習前後の、箱の位置に対する、前述の２つの中間層の出力分布を示す。特に、No.34 のニューロンは、学習前は出力分布に小さな凹凸がいくつかあったが、学習によって、右と左を区別する大域的な表現を獲得していることがわかる。

４．結論

実移動ロボットが、黒い箱を押すというタスクに対し、画像処理やタスクに関する知識を与えられることなく、視覚センサ信号を直接ニューラルネットに入力し、強化学習に基づいて学習させるだけで、適切な行動を学習させることができた。このように、従来の常識に反して、視覚センサ信号をそのままニューラルネットに入力し、タスクに関する情報も、画像処理に関する情報も一切与えずに、実際のロボットが適切な行動を学習できることを示したのは、本研究が世界で初めてであり、センサからモータまでのトータルな機能の学習という新たな方向への第一歩を踏み出したという意味で大変意義深いと考えている。しかしながら、本実験では、箱が黒で、床や周りが白い紙であるなど、理想的な実験環境で行われたものである。今後、実用性に関して鍵を握る実環境での有効性を確認することが必要である。

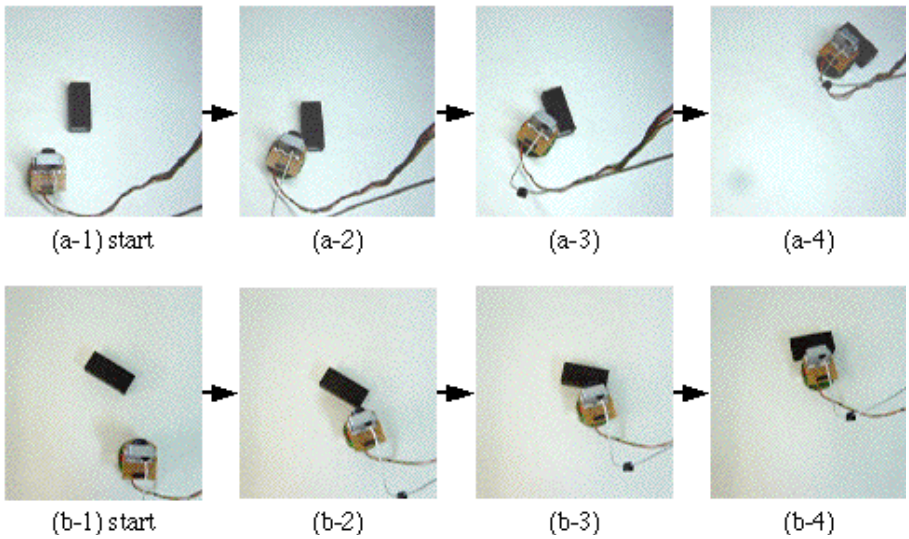


Fig. 5 Two examples of the box-pushing behaviors of the robot after learning.

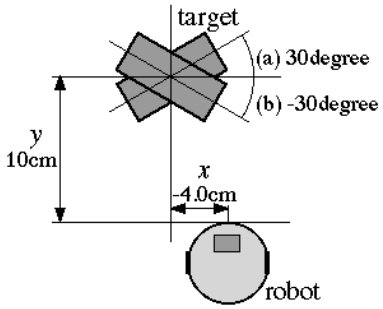


Fig. 6 Two Directions of the box employed in the following experiment.

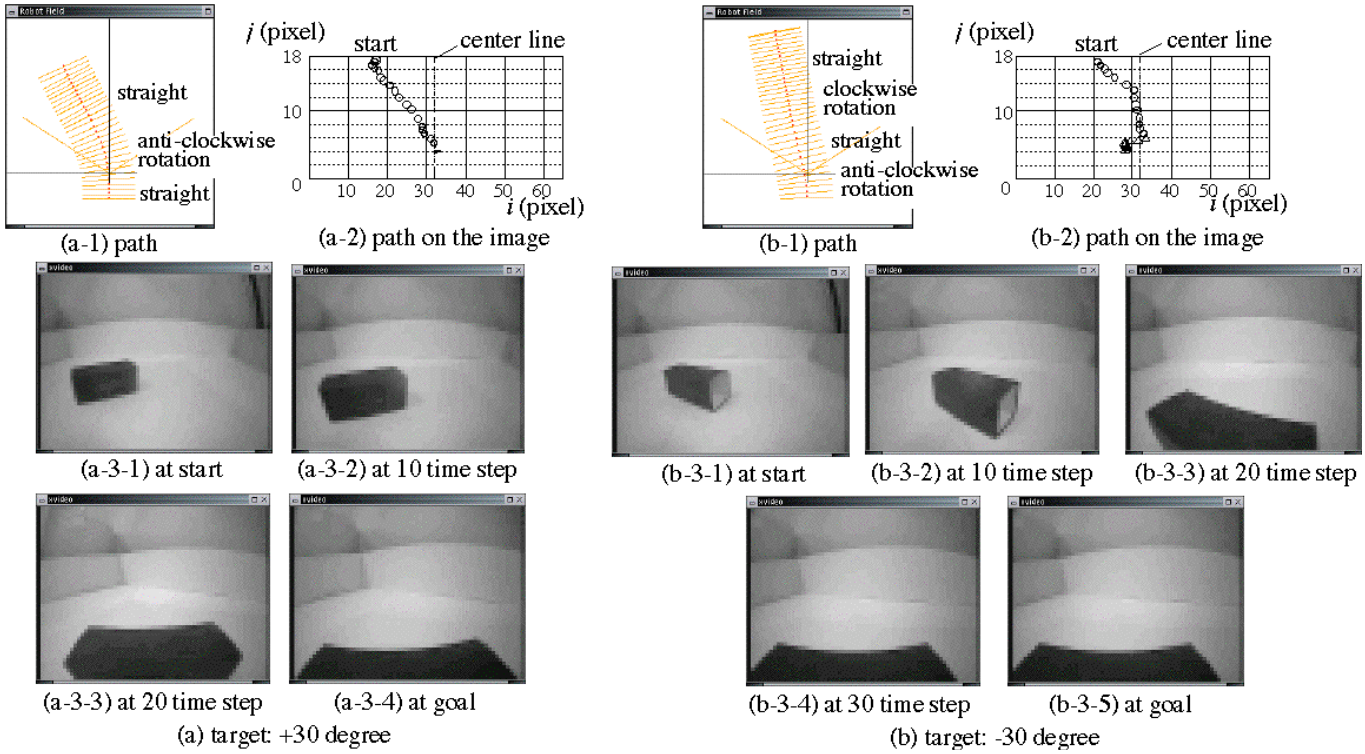


Fig. 7 The robot locus and a series of images after learning for each of the two box directions at the same initial place as shown in Fig. 6. (1) Path of the robot on the absolute coordinates (2) Path of the center of gravity of the box in the image (plot symbol indicates the number of the IR sensors that take the maximum value. ○: 0, △: 1, ●: 2) (3) A sequence of images.

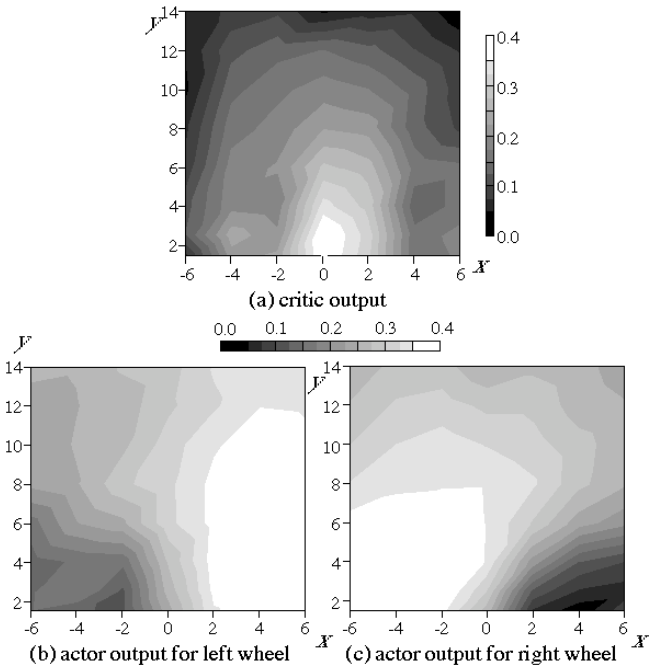


Fig. 8 The critic and actor output distribution as a function of the box location from the robot after learning.

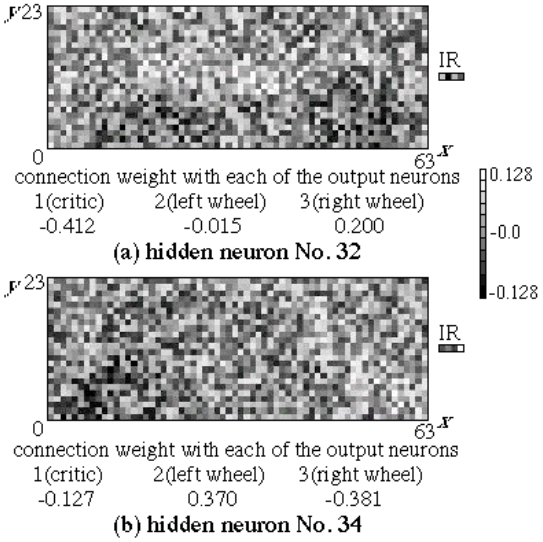


Fig. 9 The input-hidden connection weights after learning

Table 1 Correlation between weight values and the location of the input signal in the image

		before learning		after learning
(a) hidden neuron No. 32	i	-0.219	$\times 10^{-3}$ ⇒	-0.307 $\times 10^{-3}$
	j	0.085		-1.311
(b) hidden neuron No. 34	i	0.168	$\times 10^{-3}$ ⇒	0.607 $\times 10^{-3}$
	j	-0.178		-0.482

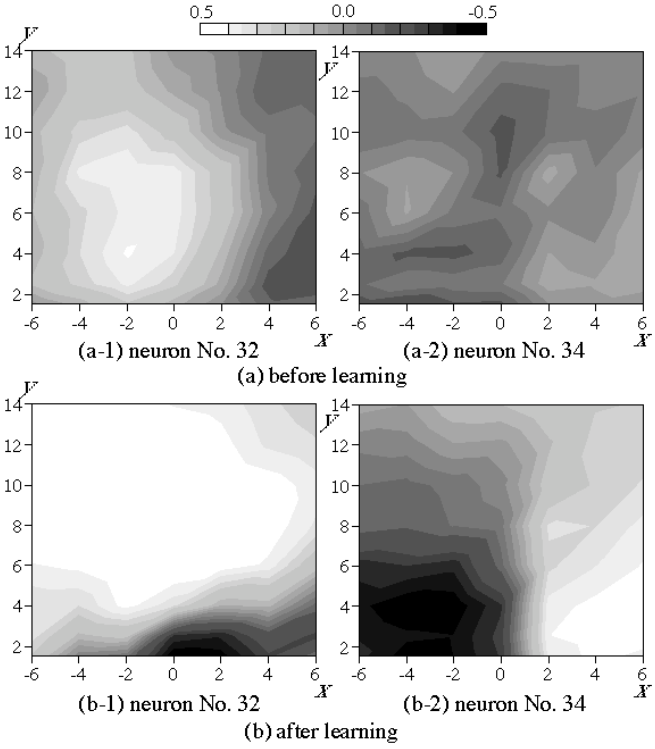


Fig. 10 Change of the hidden neuron's output distribution as a function of the box location through learning.

謝辞

本研究は、科学技術研究費補助金(#13780295, #14350227, #15300064)の補助を受けた。ここに謝意を表する。

参考文献

[1] Asada, M., Noda, S., Tawaratsumida, S. and Hosoda, K., "Purposive Behavior Acquisition for a Real Robot by Vision-Based Reinforcement Learning", *Machine Learning*, vol. 24, pp.279-303, 1996.

[2] 柴田克成, "強化学習とロボットの知能 -あめとむちで知能は作れるか？-", 第 16 回人工知能学会全国大会論文集, パネルディスカッション「強化学習とその諸相」パネリスト原稿, 2A1-05, 2002.

[3] 柴田克成, 伊藤宏司, 岡部洋一, "階層型ニューラルネットを用いたDirect-Vision-Based強化学習", *計測自動制御学会論文誌*, vol.37, No.2, pp.168-177, 2001.

[4] Iida, M., Sugisaka, M. & Shibata, K., "Direct-Vision-Based Reinforcement Learning in a Real Mobile Robot", *Artificial Life and Robotics*, vol. 7, pp. 102-106, 2004.

[5] Shibata, K. & Iida, M., "Acquisition of Box Pushing by Direct-Vision-Based Reinforcement Learning", *Proc. of SICE Annual Conf. 2003*, 0324.pdf, pp. 1378-1383, 2003.

[6] 柴田克成、伊藤宏司, "局所信号を入力としたニューラルネットにおける中間層での適応的空間再構成と汎化", *信学技報*, NC2001-153, pp.151-158, 2002.