社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
NC2000-170(2001-03)

強化学習によるリーチング動作の獲得

柴田 克成† 杉坂 政典† 伊藤 宏司††

† 大分大学工学部電気電子工学科
〒870-1192 大分市大字旦野原 700 番地
†† 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
〒226-8502 横浜市緑区長津田町 2659

E-mail: †{shibata,msugi}@cc.oita-u.ac.jp, ††ito@dis.titech.ac.jp

あらまし 本稿では、視覚センサ信号とマニピュレータの状態をニューラルネットに入力し、関節トルクを出力とし、強化学習によってリーチング動作を獲得させるシステムにおいて、(1) リンク長を可変とした場合、および、(2) 外力を付加した場合のシミュレーション結果を示す。そして、前者の中間層ニューロンの表現を観察し、入来らの道具を持たせたサルの実験において、手の到達範囲で発火するニューロンが、道具を使用することによって、道具の到達範囲で発火するようになること、また、手先とともに受容野が移動するニューロンが道具の近くでも発火するようになることと類似した結果が得られることを示す。後者では、強化学習を行うことによって、手先と環境の逆ダイナミクスを獲得できること、さらに、試行ごとにランダムな力をかけて学習させることにより、フィードバック制御に重点をおいた制御を獲得できることを示した。

キーワード 強化学習、ニューラルネット、リーチング、道具、逆ダイナミクス、フィードバック制御

Acquisition of Reaching Movement Based on Reinforcement Learning

Katsunari SHIBATA†, Masanori SUGISAKA†, and Koji ITO††

† Faculty of Engineering, Oita University
700 Dannoharu, Oita, 870-1192 Japan
†† Dept. of Computational Intelligence & Systems Science, Tokyo Institute of Technology
2659 Nagatsuta, Midori-ku, Yokohama 226-8502 Japan

E-mail: †{shibata,msugi}@cc.oita-u.ac.jp, ††ito@dis.titech.ac.jp

Abstract A learning system with a neural network whose inputs are visual sensory signals and state of a manipulator, and whose outputs are joint torques, obtains the hand reaching movement by reinforcement learning. This paper shows some simulation results in the case of (1)variable link length and in the case of (2)external force which is loaded at the hand. In the former, a hidden neuron's representation is similar to the postcentral neurons of a monkey in Iriki's experiment. The neuron activates when the food is located at the place where the hand can reach it, and the range is extended to the range where a tool can reach when the monkey uses the tool. In the latter, it is shown that the system obtains inverse dynamics of its hand and environment. When it learns in the random force at every trial, the system becomes to use feedback control mainly to control its hand.

Key words Reinforcement Learning, Neural Network, Reaching Task, Tool, Inverse Dynamics, Feedback Control

1. 従来の研究と本研究の位置づけ

1.1 強化学習

近年、報酬や罰などの強化信号という少ない情報源から、より報酬を得て罰を避ける行動を、試行錯誤をもとに、自律的に学習する強化学習が注目を集めている。強化学習は、自律的な学習であるだけでなく、合目的的、適応的な学習であり、より生物に近い柔軟な学習であると言うことができる。

また、実際に生物の中で強化学習が行われている可能性も実験によって示唆されている。Schultzらは、黒質のドーパミンニューロンが、当初は報酬に、その後次第に報酬を予測する刺激に対して反応するようになることを示した[1]。そして、ドーパミンニューロンがcriticに、その出力がTD(Temporal Difference)誤差に相当し、線条体がactorに相当すると考えることによって、actor-criticアーキテクチャの下で、TD学習に基づく強化学習[2]によく適合することが示された[3]。それ以来、大脳基底核において強化学習が行われているのではないかという考え方や、それをベースにしたモデルが提案されている[4][5]。

一方、強化学習とRBF(Radial Basis Function)を含むニューラルネットを組み合わせた研究も盛んに行われている[6][7][8]。両者の組み合わせにより、強化学習側から見ると、従来は、状態から行動へのテーブルの作成であったものが、ニューラルネットの非線形関数近似機能により、連続状態、連続行動へ容易に対応できるようになり、その幅が大きく広がった。ニューラルネットの側から見ると、教師あり学習では、逐一教師信号が必要であったものが、強化学習によって教師信号が内部生成されることで、設計者の手間が大幅に減少して、システムの自律性、適応性が大きく向上したと考えることができる。

しかしながら、従来、強化学習は、運動部分、特にプランニングのための学習ととらえられる傾向にあった。これに対し、筆者らの一部は、センサ信号を直接ニューラルネットに入力し、その出力を直接モータ(アクチュエータ)に出力し、それを強化学習に基づいて学習することにより、単にプランニングだけでなく、センサからモータまで、認識なども含めたすべての過程を、自律的、合目的的、適応的に獲得できることを主張してきた[9]。また、注意や記憶、能動認識、さらに、コミュニケーションなどの機能も獲得できる可能性をシミュレーションを通して示してきた[10][11][12]。

1.2 リーチング運動

手先のリーチング運動は、人間の運動学習の解析などの目的で広く研究されてきている。一般に、短い距離のリーチング運動を行う場合、手先の軌跡は直線となり、速度波形は、単峰性のベル型をしていると言われ、また、真横に手を伸ばした状態から、正面に手を伸ばした状態まで移動するような場合は、円軌道となることが知られている[13]。また、できるだけ速く手を動かすように指示しても、直線に近い軌道、ベル型の速度波形になることが示されている[14]。そして、これらの軌道の生成モデルとして、トルク変化最小モデルや運動指令変化最小モデルなどが提案されている[13][15]。そして、生成された軌道に追従すべく、フィードバック誤差学習[16]によって、学習初期にはあらかじめ用意されたフィードバックに、学習が進むとフィードフォーワードに重点をおいて制御されるという考え方がとられることが多い。また、筋のモデルに、フィードバック誤差学習を適用した研究も行われている[17]。

また、未知のダイナミクスを有する環境において、人間に手先のリーチング運動を行わせると、手先の軌道は大きく曲がるが、そのままリーチング運動を続けると、再び手先の軌道は直線で速度履歴がベル型に近づき、逆に、元の環境に戻すと軌道が大きく逆側に曲がることが示されている[18][19]。さらに、未知の手先負荷をかけると、筋を同時活性させて手先のスティッフネスを高くし、学習が進むと同時にその値を小さくしていると考えられる[20][21]。これは、スティッフネスを高くすることにより、未知のダイナミクスの影響を小さくしていると考えることができる。これは、フィードバックゲインを大きくすることによって、軌道の誤差を小さくしていると言い換えることも可能である。また、学習が進むと同時活性がなくなることは、外界のダイナミクスを学習することによって、フィードフォワード制御を主にして制御することにより、エネルギー消費を小さくしたり、衝突時の衝撃をやわらげていると考えることができる。筋のモデルを導入し、強化学習によってリーチング運動を獲得させる際に、同時活性の果たす学習の加速効果や、エネルギー消費を評価に組み込むことによって同時活性度が学習の進展とともに低くなることなども研究されている[22]。

一方、入來らは、道具を使ったリーチング運動をサルに行わせた際の、頭頂間溝前壁部の体性感覚と視覚とを統合していると考えられているニューロン群の活動を観察し、その一部のニューロンにおいて、

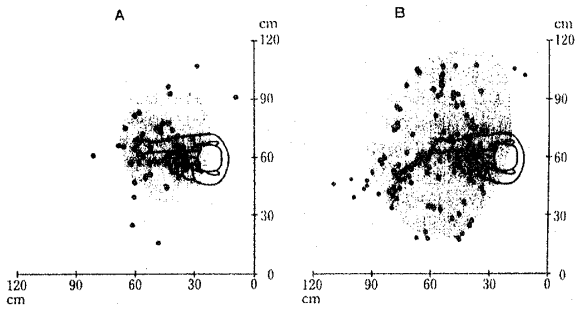

図 1 The change of the visual receptive field of a postcentral bimodal neuron by tool use. [24]

道具使用前には、手・腕の到達範囲の空間に対応した視覚受容野が、道具使用直後には、道具による到達範囲に拡大したことを示している [23] [24]。さらに、視覚受容野が手先のまわりにあり、手の動きとともに受容野も移動するニューロンの活動を観察し、手の位置が目で見えないように板で隠しても、視覚受容野が手の動きとともに移動したこと、また、サルに道具を持たせて、手と道具を直接目から見えないようにし、道具の先端の位置だけを視覚的に見えるようにすると、受容野が道具の先端に移動することなどの興味深い結果を示している [25]。そして、われわれが持つ身体イメージと関連づけて議論している。

1.3 本研究の目的

筆者らは、筋のモデルは導入していないため、生体のモデルとしては不十分な点が多いものの、すでに、視覚センサ信号と関節角、角速度の情報を入力として、手先が目標物に到達したときに与えられる報酬だけから、2 関節マニピュレータが手先のリーチング運動を実現する関節トルクを強化学習とニューラルネットの組合わせで獲得できることをシミュレーションによって示した。そして、hand-eye coordination の機能が強化学習によって獲得できること、さらに、例外があるものの、手先の軌道は概して直線に近く、速度波形がベル型に近いことを示した。[26]。また、画像上で目標物と手先が区別がつかない状態でもリーチングができること、さらに、手先の初期位置が視野の外にある場合でもリーチングが可能であり、かつ、手先が途中から視野の中に入る際も、軌道が滑らかであることを示した。

本稿では、前述の入来らの実験結果との比較を行なうため、道具の使用をリンク長の変化としてとらえ、試行ごとにリンク長を変化させ、リンク長の情報をニューラルネットの入力の一つとして追加して、手先が目標物に届かない場合も含めて学習を行なった場合の中間層の発火状態を観察する。これによって、サルの頭頂間溝前壁部のニューロンの活動を説明し、このようなニューロンが強化学習によって獲得されうることを示す。次に、文献 [18] にあるような、手先の速度に対して外部から力をかける粘性力場 (図 7) を、学習中にかけた場合と、学習後にかけた場合を比較して、強化学習によって、アームと環境のダイナミクスを獲得できること、さらに、さらに、学習中に、ランダムに力をかけることで、状況に応じて、フィードフォワードのみならず、フィードバック制御の学習も可能であることを示す。ただし、ここでは、陽な軌道は存在しないため、アームの状態が同じでも、過去の入出力関係から制御が変化することをフィードバックと呼ぶ。ニューラルネットを用いた学習によって、システムがフィードフォワードとフィードバックを適応的に組み合わせた制御を獲得できることはすでに示されているが [27]、本研究では、単に強化学習を行うだけで、フィードフォワードとフィードバックを適応的に組み合わせたハイブリッドな制御を獲得する能力があることを示す。

これらの結果を通し、生体のリーチングでも強化学習が行われている可能性を示すとともに、視覚座標系においても、関節座標系においても、軌道を陽に求める必要がないこと、さらに、フィードバックも学習によって獲得できることを主張する。軌道を陽に求める必要がない手法として、最適制御とニューラルネットを組み合わせた手法 [28] も提案されているが、本研究では、強化学習を用いることで、より自律性、適応性、柔軟性を向上させることができると考える。また、hand-eye coordination、アームや環境の逆ダイナミクス、さらには、フィードバック制御などの機能が強化学習によって獲得できることを示すことにより、筆者らが主張する、強化学習は、単なる行動のプランニングのための学習ではなく、センサからモータまでのすべてのプロセスのための学習であるという主張を後押しすることを目的とする。

2. 強化学習

ここでは、連続値動作に対応している Actor-Critic アーキテクチャ [2] を用い、Critic の学習には、時間軸スムージング学習 (TS) [29] に基づいた学習を行なう [9]。TD 学習では、Critic の出力が指数関数的に増加するように学習する一方、TS 学習では、Critic の出力が直線的に増加するように学習する点が異なる。

Critic、つまり状態評価部の学習は、

$$\Delta V_{ideal} = (V_{max} - V_{min})/N_{max} \quad (1)$$

とした。ただし、V_{max}：状態評価値の上限、ここでは 0.4、V_{min}：状態評価値の下限、ここでは -0.4。ニューラルネットの値域は、-0.5 から 0.5 とした。状態評価値の時間傾きを適応的に変化させるため、ゴールにたどり着いたところで、N_{max} を次のように学習させる。

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

ここで、$N[i]$: i 番目の試行における所要ステップ数、β: 減衰項 ($0.0 < \beta < 1.0$, ここでは、0.9996)。ただし、$N[i] > N_{max}[i-1] \times 1.5$ の場合には、$N_{max}[i]$ は変化させず、その後しばらくの間、施行中に目標物を少しずつ手先に近づけた。この状態評価値の時間傾きは、TD 学習における割引率に相当する。そして、実際の状態評価値の変化とこの理想値を比較し、1 単位時間前の状態評価値 $V(t-1)$ を

$$V_s(t-1) = V(t-1) - \eta(\Delta V_{ideal} - \Delta V(t)) \quad (3)$$

にしたがって学習する。ただし、$\Delta V(t) = V(t) - V(t-1)$、$\eta$：学習係数。手先が目標に到達するか、関節角が上限を越えた場合は、それぞれ 0.4 と -0.4 を教師信号として状態評価値を学習する。

関節トルクは、actor の出力である動作信号と試行錯誤の成分である乱数の和に比例させる。乱数は、一様乱数を 3 乗したものを用い、その乱数の値域は、状態評価値の相対的な変化量 $\Delta V/\Delta V_{ideal}$ にしたがって変化させた。動作信号 $\mathbf{m}$ は、

$$\mathbf{m}_s = \mathbf{m} + \zeta \mathbf{rnd} \Delta V \quad (4)$$

という教師信号によって学習させた。ただし、ζ: 学習定数とした。

3. 問題設定

本論文では、図 2のように、2 リンクアームを想定し、手先を視覚センサに映った目標物に到達させるというタスクを学習させる。視覚センサは、$5 \times 5 = 25$ のセンサセルを有する簡単な視覚センサを想定した。各センサセルの受容野は、正方形とし、すきまもオーバーラップもなく 2 次元状に配置されている。出力は、受容野中に手先と目標物が占める面積の割合とし、手先および目標物は、1 個のセンサセルと同じ大きさとし、両者は視覚センサ上で区別できないと仮定した。そして、ニューラルネットへは、25 個の視覚センサ信号と 2 つの関節角および角速度を入力とした。ただし、関節角、角速度ともに、一つの連

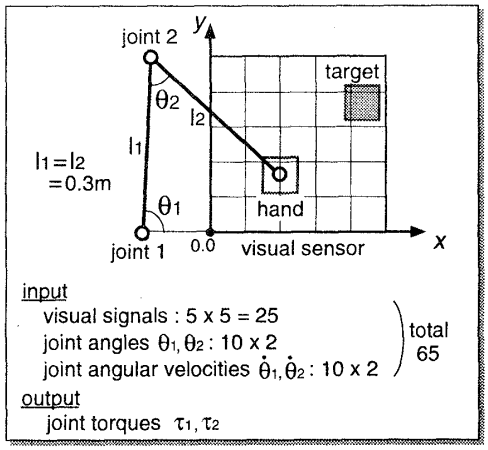

図 2 The robot hand-reaching task.

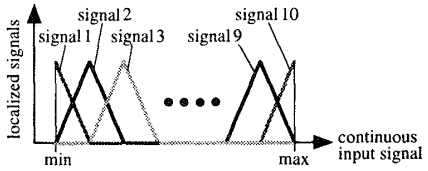

図 3 Localization of a continuous input signals.

続値信号を Fig. 3のように局所化した 10 個の信号に分解してから入力した。ただし、外力を付加したシミュレーションでは、関節角 0 を 8 個に、関節角 1 を 12 個に分割した。毎回、視野内から手先の初期位置をランダムに決定し、同じくランダムに視野内に置かれた目標物と接触し、かつ、手先の速度がある一定の値より小さい場合にのみ報酬、つまり、0.4 を教師信号として学習した。また、どちらかの関節角度が 0 度より小さくなるか、関節角 0 が 90 度より大きくなった場合には、そこで試行を中断し、critic を-0.4 を教師信号として学習した。それ以外の場合は、強化信号は 0 とした。ニューラルネットは中間層 2 層の 4 層とし、中間層のニューロン数は、下位から 30 個、10 個、出力層は、critic 1 個と各関節トルクに相当する actor 2 個の計 3 個とした。

アームのダイナミクスは、以下の通りとした。

$$\begin{aligned} \tau_1 &= (I_1 + I_2 + 2M_2 l_1 s_2 cos\theta_2 + M_2 (l_1)^2 \ddot{\theta}_1 \\ &\quad + (I_2 + M_2 l_1 s_2 cos\theta_2)\ddot{\theta}_2 \\ &\quad - M_2 l_1 s_2 (2\dot{\theta}_1 + \dot{\theta}_2)\dot{\theta}_2 sin\theta_2 + B_1 \dot{\theta}_1 \quad (5) \\ \tau_2 &= (I_2 + M_2 l_1 s_2 cos\theta_2)\ddot{\theta}_1 + I_2 \ddot{\theta}_2 \\ &\quad + M_2 l_1 s_2 (\dot{\theta}_1)^2 sin\theta_2 + B_2 \dot{\theta}_2 \quad (6) \end{aligned}$$

ただし、M_i, l_i, s_i と I_i は、それぞれ、リンク i の質量、長さ、ジョイントから質量中心までの距離、関節回りの慣性モーメントを表す。ただし、関節角 1 が

表1 Parameters used in the dynamical arm model.

Parameter	link1	link2
Mi (kg)	2.0	
li (m)	0.3	
si (m)	$li/2$	
Ii $(kg\ m^2)$	$Mi * li^2 / 3.0$	
Bi $(kg\ m^2/s)$	0.4	0.2
$\tau_{max\,i}$ $(N\ m)$	4.0 or 5.0	2.0 or 3.0

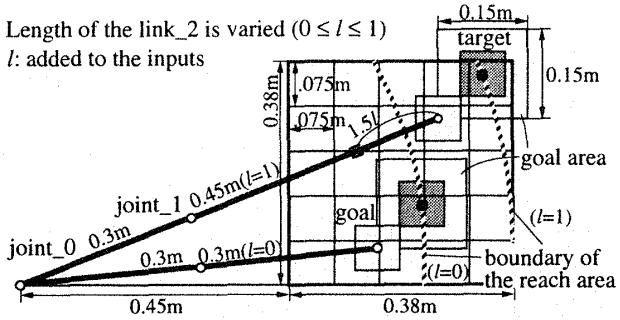

図4 The task setting in the case of variable link length.

180度より大きくなった場合は、関節角1を180度に固定した一つのリンクとして計算した。各パラメータは、表1のように設定した。トルクの最大値は、リンク長可変のシミュレーションの際には4.0[Nm]と2.0[Nm]を、外力を付加したシミュレーションでは、5.0[Nm]と3.0[Nm]とした。上記のアームの運動方程式は、ルンゲクッタ法で、0.02秒のサンプリング間隔で、数値積分を行なって解き、強化学習も同様の時間間隔で適用した。

4. シミュレーション

4.1 リンク長可変の場合

まず始めに、道具の使用をリンク長の変化ととらえて、第2リンクの長さ l_2 を0.3mから0.45mの間の連続値として、各試行ごとに乱数を用いて変化させて学習を行なった。この長さの変化を0から1の間に正規化し (l)、これを図3のように局所化して10個の信号に変換して入力した。このとき、アームを図4のように配置し、第2リンクが一番短いと視野の右の方にある目標物には手先が届かず、一番長い場合でも、視野の右上の端の方にある目標物には手先が届かないようにした。図中の2つの破線が、リンク長が一番短い場合と一番長い場合の、手先が届く目標物の中心位置の範囲の境界をそれぞれ表している。また、ここでは、視覚センサの解像度が十分でないため、学習を安定させるために、境界から0.015mの距離には目標物を置かなかった。

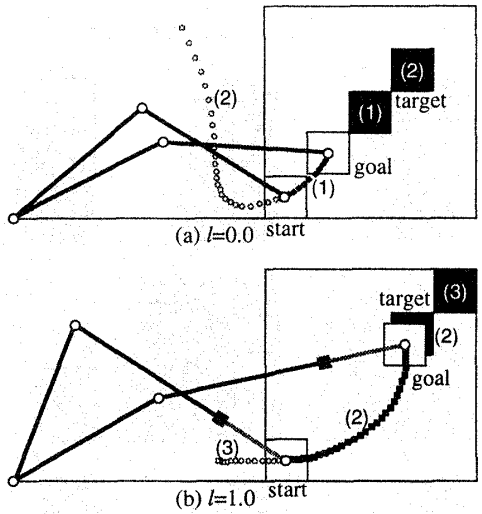

図5 The difference of the hand trajectories depending on the link length.

学習後の手先の軌道の例を図5に示す。第2リンクが短い場合も長い場合も、目標物が届く範囲にある場合は手先を目標物の方に伸ばし、届かない場合は、目標物とは関係のない方向に手先が動いていることがわかる。また、目標物が同じ場所(2)でも、第2リンクの長さによって、届く場合にのみ手先を目標物に向かって伸ばしていることがわかる。前述の入來らの報告の中で、サルは、最初に見て、道具を使って届く範囲にエサがあれば取るが、届く範囲になければ、最初からエサとりの行動を起こさないとしている[24]。ここでは、届かない場合、目標物と関係のない方向に手先が動いているが、明らかに、届く場合とは違う行動を起こしている。また、エネルギー消費などの評価指標を導入すれば、届かない場合の無意味な行動を抑制できる可能性がある。

図6に、目標物の位置に対するcriticの出力、および上位の中間層の2個のニューロンの値の分布を、第2リンクが最も短い場合と長い場合について、また、それぞれ、左下、左上の2つの手先位置に対して示す。関節角速度は0.0とした。ただし、中間層ニューロン1の出力値は、正負を反転して示した。これより、criticの出力は、破線で示した手先の到達境界を境に、到達できないところは小さい値に、到達できるところは、手先に近いほど大きな値になる傾向がある。また、中間層ニューロン1の出力は、到達境界で、criticの出力より値が大きく変化しており、かつ、手先の位置にあまり影響を受けていないことから、手先が目標物に到達できるかどうかを表現していると考えることができる。これは、ちょうど、入來らのサルによる実験結果(図1)と一致する。

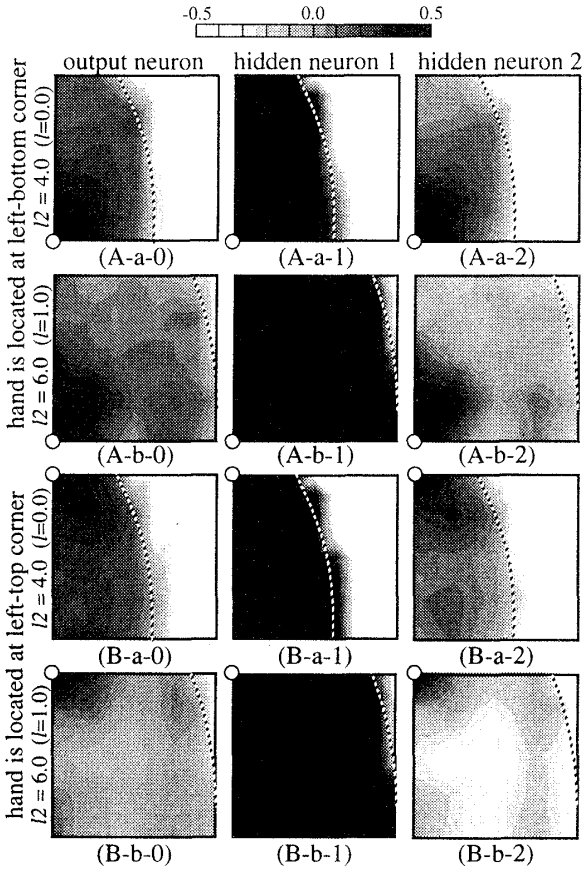

図6 The distribution of the neurons' outputs as a function of the target location.

上位の中間層ニューロン10個のうち、値の正負が逆のものを含め、このような特性のニューロンが4個あった。入來らの報告では、これらのニューロンは、肩などの体性感覚にも反応するとしているが、これに相当する結果を示すことはできなかった。しかし、身体像をコードするニューロンが、強化学習によって獲得できる可能性を示せたと考える。

また、中間層ニューロン2の出力は、ciritic の出力と似ているが、より目標物が手先の位置に近いほど大きな値を示している。これにより、入來らの報告における、手のまわりに受容野が形成され、手とともにその受容野を移動するニューロンが、道具を使わせると道具まで受容野が大きくなること、また、道具の先端のみを表示させるとそこに受容野が形成されることを説明することができる。また、手先を見せないで目標物のみ見せて学習をさせれば、手先が見えなくても手先が存在する周辺に受容野が形成されることも予測される。ニューラルネットの初期値を変えて行なった別のシミュレーションでは、中間層ニューロン1のタイプのニューロンはやはり4

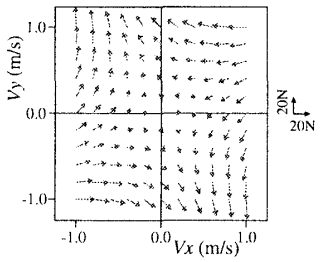

図7 The force field loaded to the hand.

個あったが、中間層ニューロン2のように、手先の近傍のみ大きな値をとるニューロンは存在しなかった。

このようなニューロンができる理由を考えてみる。まず、図中の critic の出力、つまり状態評価値の分布は、手先が到達しない状態では、いくらどのような行動を起こしても報酬をもらうことはあり得ないので、式(3)によって、値が小さくなるが、到達できる状態では、到達することによって、そこに至るまでの状態評価値が大きくなる。したがって、図のように、評価値が境界部分で大きく変化するようになる。さらに、到達できる範囲内でも、到達するまでの時間が短いほど大きな値に出力を連続的に変化させる必要がある。そして、中間層ニューロン1は、そのうち、到達範囲のみをコーディングし、中間層ニューロン2は、到達できるまでの時間を主にコーディングし、その結果、手先に近い部分が発火すると解釈することができる。また、中間層ニューロン1は、同じような出力特性を示すものが10個のうち4個もあったが、これは、関節角速度が0の場合だけを観察しているため、関節角速度が値を持つ場合には、それぞれ違う働きをしている可能性もある。

4.2 外力を付加した場合

次に、手先に外力を付加した場合のシミュレーションを行なう。(1) 文献[18]にあるような、手先の速度に応じて外部から力をかける粘性力場(図7)をかけた場合と(2)10試行中に3試行の割合で、x, y 方向それぞれ絶対値が8.0[N]以下の範囲でランダムに決定した外力を試行中に付加し続けた場合および(3)外力を全く付加しなかった場合についてリーチングの学習を行なった。タスクの設定を図8に示す。ここでは、アームの長さは固定とした。また、必要に応じてフィードバックを学習できるように、1単位時間前の関節角、角速度、およびトルクの値を局所化してニューラルネットに入力として与えた場合とそれらを与えなかった場合の学習を行なった。

過去の状態入力を入れた場合について、上記(1),(2),(3)のそれぞれの学習を行なったものに対し、

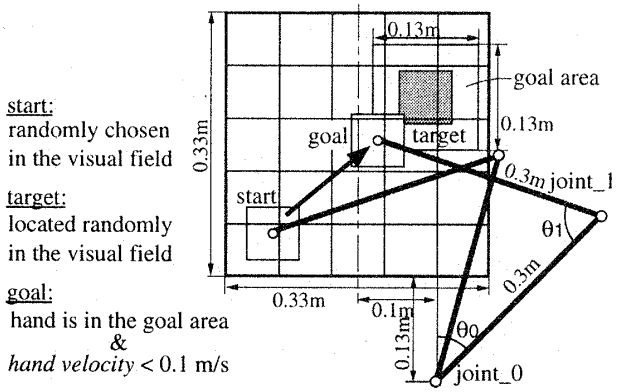

図 8 The task setting in the case of the external force load.

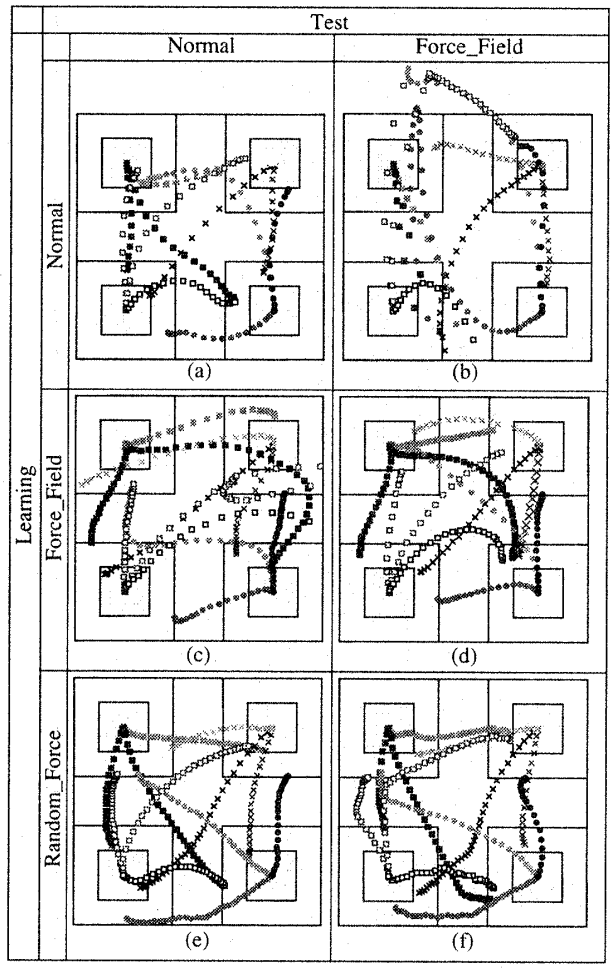

図 9 The examples of the trajectories.

外力を付加しなかった場合と粘性力場を付加した場合の軌道の例を図 9に示す。目標物の位置と手先の初期位置を、図中の右上、右下、左上、左下の 4ヵ所から選んだ 12 の組み合わせについて示した。これより、学習時、テスト時ともに粘性力場をかけなかった場合と、ともにかけた場合の軌道は比較的直線に近く、テスト時のみかけたものは、粘性力場の影響

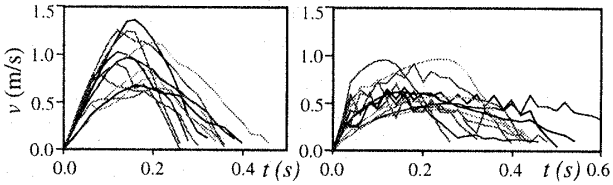

図 10 The examples of velocity shape.

表 2 Learning results. The 3 numbers indicate the number of success-out-fail respectively. The number in parenthesis indicates the average time.

learning \ test	no force field	force field
no force_field	2071- 5- 4 (0.324)	1139-926- 15 (2.03)
force_field	1869-193- 18 (0.793)	2035- 38- 7 (0.467)
random_force	2038- 5- 37 (0.546)	1833-227- 20 (0.909)
random_force (no past state inputs)	1887- 0-193 (0.896)	1165-853- 62 (2.12)

を受けた軌道に、学習時のみかけたものは、粘性力場と逆の方向に影響を受けた軌道になっているが、テスト時のみかけたものより影響が小さいことがわかる。また、ランダムに外力をかけた場合は、力場をかけた場合とかけなかった場合で軌道の差が小さい。学習時、テスト時ともに力場をかけなかった場合と、ともにかけた場合の手先速度の変化を図 10に示す。これより、力場をかけなかった場合は、比較的ベル型に近い速度波形をしているが、力場をかけた場合は、あまりベル型に近くないことがわかる。

また、目標物と手先を、それぞれ 8 × 8 の格子上 64ヵ所に順番に置き、最初から到達している場合を除いた 1040 の組み合わせに対し、それぞれの成功回数、関節角のリミットを越えた回数、4 秒以内に到達しなかった回数をニューラルネットの初期値を変えた 2 回のシミュレーションの結果を足したもの、および失敗時を 4 秒とした平均到達時間を表 2に示す。これらの図表より、力場をかけなかったときもかけた時も、それぞれの逆ダイナミクスを学習しており、環境のダイナミクスが変わるとうまくリーチングができなくなっていることがわかる。また、ランダムに力をかけた場合は、テスト時に力場をかけたときもかけなかった場合も、いずれも比較的成功回数が多い。ところが、ニューラルネットの入力から過去の状態と出力を抜いて学習させると、力場の場合にうまくリーチングができなくなっている。ランダムな外力の場合は、試行を始めるまでは、どのような力がかかるかを予測することができないため、現在

の状態のみによる逆ダイナミクスに基づくフィードフォワード制御では対応することができず、制御の重点がフィードバックに移った方がより良い制御ができると考えられる。本システムは、強化学習を通して、フィードバックに重点を置いた制御をするように、経験から適応的に学習を行ったと考えられる。

5. おわりに

可変リンク長のアームを用いたリーチングタスクを、強化学習によって、視覚センサ信号と関節角、角速度を入力とし、トルクを出力とするニューラルネットで学習したところ、中間層に、手先の届く範囲を表現するニューロン、および、手先の周辺に受容野を持つニューロンが見つかった。また、粘性力場を付加してリーチングタスクを学習させることにより、ニューラルネット内にアームと環境の逆ダイナミクスを獲得し、フィードフォワード制御ができようになるとともに、試行ごとにランダムな外力を付加した環境では、フィードバックに重点をおいた制御を学習によって獲得できることがわかった。

本研究での結果を生体と比較する場合、トルク制御と筋制御という決定的な違いがあるため、説得力に欠けることは否めない。今後、早急に筋のモデルを導入して比較していきたい。

謝　辞

本研究の一部は、文部省科学研究費基盤研究(#10450165)、および日本学術振興会未来開拓研究プロジェクト“生物的適応システム”(JSPS-RFTF96 I00105)の援助による。ここに、謝意を表する。

文　献

[1] Schultz, W., et al., “Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task”, *J. Neurosci*, **13**, pp. 900–913, 1993.

[2] Barto, A.G., et al., “Neuronlike adaptive elements that can solve difficult learning control problems,” *IEEE Trans. of SMC*, **13**, pp. 835–846, 1983.

[3] Houk, J.C., et al., “A model of how the basal ganglia generate and use neural signals that predict reinforcement”, *Models of Information Processing in the Basal Ganglia*, pp. 249–270, MIT Press, 1995.

[4] Doya, K., “What are the computations of the cerebellum”, *Neural Network*, **12**, pp. 961–974, 1999.

[5] Nakahara, H., et al., “Reinforcement learning with multiple representations in the basal ganglia loops for sequential motor control”, *Proc. of IJCNN'98*, pp. 1553–1558, 1998.

[6] Anderson, C.W., “Learning to control an inverted pendulum using neural networks”, *IEEE Control System Magazine*, **9**, 31-37, 1989.

[7] Tesauro, G.J., “Practical issues in temporal difference learning”, *Machine Learning*, **8**, pp.257–277, 1992.

[8] 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”, 電子情報通信学会論文誌, **J82-D-II** (11), pp. 2118–2131 (1999)

[9] 柴田克成, 岡部洋一, 伊藤宏司, “Direct-Vision-Based 強化学習 －センサからモータまで－”, 計測自動制御学会論文誌, **37** (2), 2001.

[10] 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, pp. 153–160, 2000.

[11] Shibata, K., Nishino, T. and Okabe, Y., “Actor-Q based active perception learning system”, *Proc. of ICRA2001 (IEEE Int'l Conf. on Robotics and Automation)*, 2001. (to appear)

[12] 柴田克成, 伊藤宏司, “利害の衝突回避のための交渉コミュニケーションの学習と個性の発現–リカレントニューラルネットを用いたダイナミックコミュニケーションの学習–”, 計測自動制御学会論文誌, **35** (11), pp. 1346–1354, 1999.

[13] Uno, Y., et al., “Formation and control of optimal trajectory in human multijoint arm movement”, *Biol. Cybern.*, **61**, pp. 89–101, 1989.

[14] 鈴木邦典, 宇野洋二, “最短時間到達運動において脳が適用する滑らかさの規範”, 信学論, **J83-D-II**, pp.711–722, 2000.

[15] Nakano, E., et al., “Quantitative examinations of internal representations for arm trajectory planning”, *J. Neurophysiology*, **81** (5), pp. 2140–2155, 1999.

[16] Kawato, M., et al., “A hierarchical neural-network model for control and learning of voluntary movement”, *Biol. Cybern.*, **57**, pp. 169–185, 1987.

[17] 片山正純, 川人光男, “筋肉・骨格系の運動制御を行なう並列階層制御神経回路モデル”, 信学論, **J73-D-II**, pp.1328–1335, 1990.

[18] Shadmher, R., “Adaptive representation of dynamics during learning of a motor task”, *J. NeuroSci.*, **15** (5), pp. 3208–3224, 1994.

[19] 窪寺裕之, 伊藤宏司, “粘性力場における上肢運動の運動学習とダイナミクスに依存した汎化”, 信学技報, **98** (673), NC98-116, pp. 137–144, 1999.

[20] 大須理英子, 道免和久ら, “運動学習時における筋活性の変化”, 信学技報, **NC96**-139, pp. 201-208, 1997.

[21] Thoroughman, K.A. and Shadmher, R., “Electromyographic correlates of learning an internal model of reaching movement”, *J. Neurosci.*, **19** (19), pp.8573-8588, 1999.

[22] 井澤淳, 近藤敏之, 伊藤宏司, “強化学習法を適用した人腕運動学習制御における粘弾性調節戦略”, 第13回自律分散システムシンポ資料, pp. 123–128, 2001.

[23] Iriki, A., et al., “Coding of modified body schema during tool use by macaque postcentral neurons”, *Neuroreport*, **7**, pp. 2323–2330, 1996.

[24] 入來篤史, “サルの道具使用と身体像”, 神経進歩, **42** (1), pp. 98–105, 1998.

[25] 入來篤史, “道具を使う手と脳の働き”, 日本ロボット学会誌, **18** (6), pp.786–791, 2000.

[26] Shibata, K., Sugisaka, M. and Ito, K., “Hand reaching movement acquired through reinforcement learning”, *Proc. of The 2000 KACC (Korea Automatic Control Conf.)*, 90rd (CD-ROM), 2000.

[27] 柴田克成, 稲葉雅幸, 井上博允, “ニューラルネットによるロボットの運動学習”, 第6回日本ロボット学会学術講演会予稿集, pp.141–142, 1988.

[28] Zheng, X.-Z., Inamura, W., Shibata, K., and Ito K., “A learning and dynamic pattern generating architecture for skillful robotic baseball batting system”, *Proc. of ICRA2000*, pp. 3227–3232, 2000.

[29] 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.