Direct-Vision-Based 強化学習に基づく Hand-Eye Coordination の形成

○柴田 克成, 伊藤 宏司
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

Formation of Hand-Eye Coordination Based on Direct-Vision-Based Reinforcement Learning

Katsunari SHIBATA & Koji ITO
Dept. of Computational Intelligence & Systems Science, Tokyo Institute of Technology
shibata@ito.dis.titech.ac.jp

Abstract: This paper shows that a robot with hand-eye system can learn hand-reaching tasks by neuro-based reinforcement learning in which a reward is given only when the hand reaches the target. This system consists of a neural network whose inputs are raw visual sensory signals, joint angles of the arm, and the binary value indicating the existence of the obstacle. The target, obstacle, and hand locations are not calculated explicitely. By the analysis of the hidden neurons' representation after the learning, it was known that the target location was not represented independently from the hand location. Furthermore, the representation of the hand location in the hidden layer is acquired by integrating the joint angles and visual signals.

1. はじめに

近年、自律学習能力から強化学習が注目されている。これをロボットの制御に用いる際には、通常、まずセンサ信号からいくつかの状態へ分類し、その各状態から適切な行動へのマッピングを強化学習によって学習する[1)]。この状態分け、つまり、状態空間の設計が難しい問題となっている。一方、状態-行動間のマッピングを非線形連続値とするために、ニューラルネットも使われている[2)]。

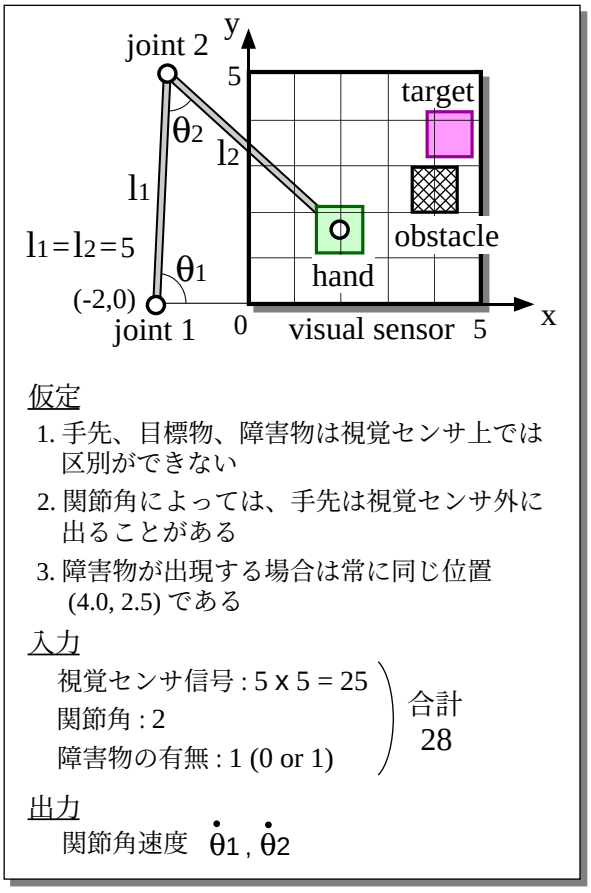
著者らは、局所的な受容野を持つセンサセルからの信号をニューラルネットに直接入力して強化学習を行う Direct-Vision-Based 強化学習を提案してきた。これによって、中間層に連続かつ適応的な状態空間が形成されるとともに[3)]、学習が高速化することを示した[4)]。ニューラルネットを用いてシステム全体をシームレスな構造とすることにより、強化学習が単に行動のプランニングとしてだけでなく、認識やコミュニケーション等の高次の機能にどこまで有効であるかを検証している。本稿では、Hand-Eye Coordination、つまり、生の視覚情報と関節角情報との関係の把握およびそれをいかに統合するかに着目する。そして、タスクとして、**図 1** に示すようなリーチングタスクを取りあげ、強化学習で異種センサの情報が統合され、制御が適切に行われるかどうかを調べる。そして、学習後に、ニューラルネットの中間層が手先や目標物の位置をどのように表現しているかを調べる。

われわれ自身を振り返ってみると、何かの物体に自分の手を伸す際に、その手を意識することはほとんどない。しかし、自分の手を見せずに育てたサルに、生後 34 日後に初めて自分の手を見せると、あたかも他の物体を見るようにしげしげ眺めたという報告がある[5)]。この実験は、Hand-Eye Coordination が生後獲得されるものであることを示唆している。また、経験を積むことによって自分の手が意識されなくなるということは、われわれやサルが、手の位置を陽に計算して、画像から手のイメージを取り除き、物体の位置を求めているのではなく、視覚から運動への直接的なマッピングが存在しているのではないかと推測される。

工学的に視覚フィードバックを実現するための Hand-Eye Coordination に関する研究は多数なされている。中でも Jägersand らは、視覚情報からロボットの関節角への変換を、モデルや学習のための特別なステップを必要とせずに学習する方法を示している[6)]。しかし、ここでは、視覚イメージから目標物の位置を抽出できることが前提となっている。したがって、単に視覚座標と関節角座標の間の変換 (visual motor Jacobian) を経験から学習しているということになる。

本稿で取り上げたタスクは、冗長自由度もなく非常に単純なものであるが、目標物、障害物、手先の位置を陽に計算することなく、Hand-Eye システムがリーチングを強化学習によって学習できることを示す。また、ここでは、本稿の手法の有効性を示すため、次の 2 つの仮定を設けた。1) 視覚イメージ中で、目標物、障害物、手先が区別できない。2) 関節角によっては手先が画像から消えることがある。前述のように、視覚フィードバックをかけるためには、通常、手先、目標物、障害物の位置

キーワード：Direct-Vision-Based Reinforcement Learning, Neural Network, Hand-Eye Coordination, Reaching Task, Hidden Representation



仮定

1. 手先、目標物、障害物は視覚センサ上では区別ができない
2. 関節角によっては、手先は視覚センサ外に出ることがある
3. 障害物が出現する場合は常に同じ位置 (4.0, 2.5) である

入力

視覚センサ信号 : 5 x 5 = 25
関節角 : 2
障害物の有無 : 1 (0 or 1)
合計 28

出力

関節角速度 $\dot{\theta}_1, \dot{\theta}_2$

図 1 例題として用いたリーチングタスク

をそれぞれ計算する。しかし、ここでは、視覚センサ情報の前処理を行おうとしても、3 つの物体の区別がつかない上、手先が視覚センサ上にない場合もある。もちろん、画像上の手先位置とアームの関節角の関係および障害物の位置が常に同じであることが予めわかっていれば目標物の位置を知ることができる。しかし、強化学習は、タスクに関する情報が不十分なときにその威力を発揮する。したがって、前処理が不要であることは、強化学習の利点を最大限引き出すことにつながると考えられる。

2. タスク設定

図 1 に本稿で用いたタスクを示す。視覚センサは、$5 \times 5 = 25$ 個の視覚センサセルからなり、それぞれのセルは、受容野中の目標物、障害物、手先が占める面積の割合を 0 から 1 の連続値で出力するものとする。各センサセルの大きさは、1×1 とし、目標物、障害物、手先の大きさも同じとした。視覚センサの左下隅を原点として、アームの付け根は (-2.0, 0.0) に固定し、手先の初期位置は、$-1.99 \leq x \leq 6.01$, $-1.99 \leq y \leq 6.01$、かつ腕が届く範囲内でランダムに決定した。目標物の位置は、視覚センサがとらえられる範囲内、つまり、$0.5 \leq x \leq 4.5$, $0.5 \leq y \leq 4.5$ 内で乱数で決定した。障害物は、乱数を使って 1/2 の確率で出現し、場所は常に (4.0, 2.5) とした。そして、目標物と障害物は、1 試行の間動かないとした。また、障害物がないときは、目標物や手先がその位置に来ることもある。アームの各リンクの長さは 5 とし、関節角は、0.0 から π の間に制限し、それを越えないものとした。関節角と手先の位置は、(-2.0, 0.0) を除いて 1 対 1 の関係にあり、2 つの関節角は観測できるものとした。ロボットへの入力は、25 個の視覚信号と 2 つの関節角、それから障害物の有無を示す 0 または 1 の信号の計 28 個である。出力は、2 つの関節角速度である。そして、手先が目標物に触れると報酬がもらえるものとした。また、手先が障害物に触れた場合および関節角が制限を越えようとした場合は小さな罰を与えた。

3. 画像からの情報抽出の学習

視覚センサ信号とその他の信号の両方を入力として使ったときのニューラルネットの基本的な学習能力を調べるため、画像上の目標物位置の抽出を教師あり学習で行わせた。設定はほとんど前節で述べた通りであるが、ニューラルネットの出力の値域は -0.5 から 0.5 とし、目標物の位置 x と y を -0.4 から 0.4 の間に正規化して教師信号として与え、Back Propagation で学習した。手先と目標物の位置は、毎回ランダムに決定する。中間層のニューロン数は 20 とした。目標物と手先は重なることもあるとした。一方、障害物は、手先や目標物と重ならないとした。

図 2 に、学習後の目標物の位置 x, y と出力の関係を示す。それぞれのグラフは、ランダムに抽出した 100 回分の出力をプロットした。手先の位置や障害物の有無がランダムに変化し、時には手先が画像から消え、そして、時には、目標物が手と重なるにもかかわらず、出力は直線で表された教師信号とほぼ一致していることがわかる。このことから、ニューラルネットは、生の視覚センサ情報と関節角の情報から目標物の位置を抽出する能力を持っていることがわかる。

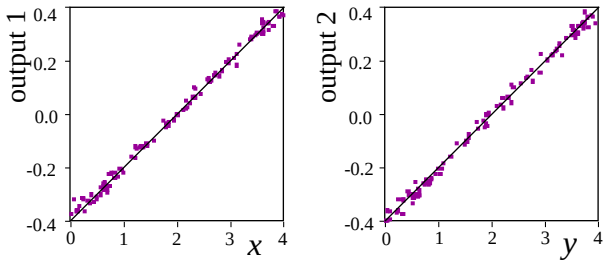


図 2 教師あり学習による視覚センサ信号から目標物の位置抽出の学習結果

4. 強化学習

この節では、強化学習のアーキテクチャとアルゴリズムを紹介する。アーキテクチャは Actor-Critic[7] をベースとするが、両者を 1 つの階層型ニューラルネットで構成し、状態評価と動作信号の両方の出力を持つ。(**図 3** 参照) これは、共有できる情報を中間層で共有するとともに、共有できない情報も必要に応じて中間層ニューロンを両者が柔軟に使い分けることを期待している。

アルゴリズムは、時間軸スムージング学習 (TS) に基づく強化学習アルゴリズムである[8]。これは、Temporal Difference(TD) ベースの強化学習[7]と似ており、状態評価値の理想時間変化が指数関数の代わりに直線となったものである。つまり、TS-based の学習では、タスク達成のための所要時間が状態評価値として表現される。

まず、状態評価値の理想時間変化量 ΔV_{ideal} を過去のタスク達成に要した時間の最大値 N_{max} から

$$\Delta V_{ideal} = V_{amp}/N_{max} \tag{1}$$

と求める。ただし、V_{amp} : 状態評価値の理想値域で、ここでは 0.8 (最大値 0.4、最小値 -0.4)。また、N_{max} は、学習の進行にあわせて

$$N_{max}(t) = \begin{cases} N(t) & \text{if } N(t) > \beta N_{max}(t-1) \\ \beta N_{max}(t-1) & \text{otherwise} \end{cases} \tag{2}$$

とし、過去のことは少しずつ忘却させる。ただし、$N[t]$: 時刻 t の試行に要した時間、β : 忘却の係数 $(0.0 < \beta < 1.0)$ である。ここでは、ニューロンの出力値は、-0.5 から 0.5 とした。そして、実際の出力値を理想値と比べることにより、一単位時間前の状態評価値 $V(t-1)$ は、

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

という教師信号で学習される。ただし、V_s : 状態評価値の教師信号$\Delta V(t) = V(t) - V(t-1)$, η : 学習係数である。この学習により、状態評価値は、時間に対して滑らかに変化し、その傾きは一定値に近づく。この傾きは、TD-based の強化学習の割引率に相当する。そして、ロボットが目標状態に達したときに、状態評価値は 0.4 という教師信号で学習される。

動作は、ニューラルネットの出力である動作信号 $\mathbf{m}$ と試行錯誤の成分である乱数 $\mathbf{rnd}$ の和に基づいて行う。動作信号 $\mathbf{m}$ は、

$$\mathbf{m}_s = \mathbf{m} + \zeta \mathbf{rnd} \Delta V \tag{4}$$

という教師信号によって学習する。ただし、ζ : 学習係数である。これによって、状態評価値の変化がより大きくなるような動作を学習していく。この動作の学習は、状態評価値の学習と並列に行う。

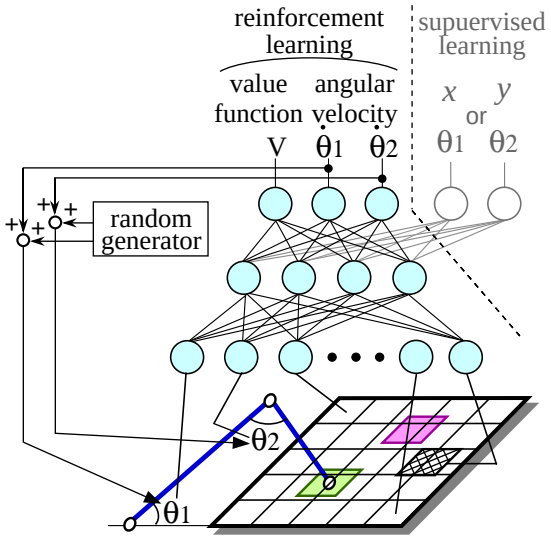


図 3 リーチングを学習するロボットのシステム構成。出力の右の 2 つは、中間層の表現を解析するために使用。

5. シミュレーション

5.1 リーチングの学習

図 3 に、ここで用いたニューラルネットの構成を示す。このうち、出力層の右側の 2 つのニューロンは、中間層の解析用であり、強化学習時には用いない。ニューラルネットは、連続値の視覚センサ信号、同じく連続値のアームの関節角、および 2 値の障害物の有無の信号を入力として受け取る。中間層の数は 1 層で、ニューロン数は 40 とした。動作信号出力は 2 個で、微小な乱数 (一様乱数を 3 乗したもの) を加え、0.1 をかけてそれぞれの関節の角速度とした。また、アームは誤差なく追従できるものとした。乱数の振幅は、$\Delta V_{ideal}/\Delta V$ に基づいて、状態評価値のゲインが小さいときには振幅が大きくなるように適応的に変化させた。ニューラルネットの出力の値域は、-0.5 から 0.5 とし、式 (3) と 式 (4) で求められる教師信号は、-0.4 から 0.4 の範囲に制限した。以上より、最大関節角速度は 0.04 であり、90 度回転するのに必要な最小時間は 32 である。もし、ある一定時間内に手先が目標物まで到達しなかった場合は、それ以後の試行で、目標物を徐々に手先に近づけてやる。また、関節角が 0.0 より小さくなったり、π より大きくなった場合は、0.0 または π とし、0.1 という小さな罰を与えた。また、手先が障害物と衝突する動作信号を出力した場合は、手先を動かさず、やはり 0.1 の罰を与えた。

図 4 に、400000 試行の学習後の手先の軌道を目標物位置と障害物の有無を変えた 2 つの場合について示した。ロボットは、障害物がある場合は、障害物を回避しながら、最終的に目標物にたどり着いていることがわか

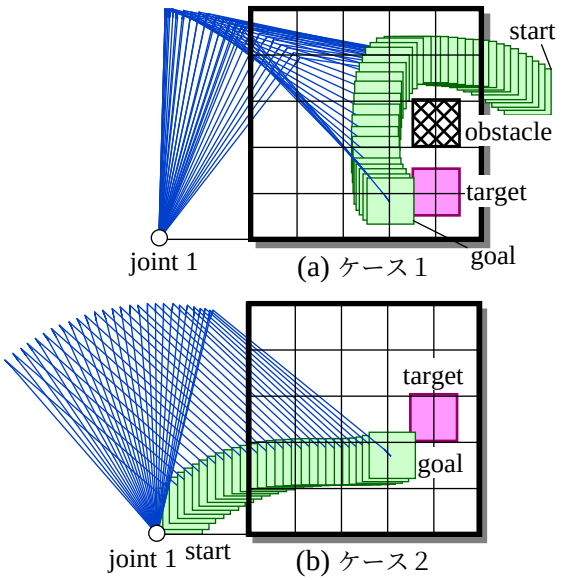


図 4 目標物の位置と障害物の有無を変えた 2 つの場合における学習後の手先の軌跡

る。また、手先、目標物、障害物の区別がつかないにもかかわらず、ロボットは自分の手先を認識し、さらに、目標物と障害物も区別できているように見える。また、初期位置で手先が視覚センサ外にあり、途中で視覚センサ上に見えるようになる場合でも、手先の軌道は滑らかに変化していることがわかる。

目標物位置と障害物の有無を**図 4** のようにした際のそれぞれの場合について、手先位置に対する状態評価関数と手先の動作ベクトルおよび**図 4** の場合の手先の軌道を**図 5** に示す。障害物があると、特に右上の角で状態評価関数に大きな変化があることがわかる。**図 5** (b) の右どなりの図に、ケース 2 の場合の手先が目標物まで到着するのに必要な最小時間を示した。この図と状態評価値が似ていることから、状態評価値は最小時間をおおむね表現できていることがわかる。しかしながら、障害物が現れる場所の近くを通る場合、障害物が存在しない場合でも、存在しない障害物を避けるような傾向が少し観察された。ただし、障害物があるときの動作とは明らかに違っていた。さらに、手先が初期位置の範囲の端に置かれた場合、手先が障害物にぶつかったり、関節角の制限を越える場合があった。しかし、試行錯誤成分により少ない回数でそのような状態から逃れることができた。これは、ロボットがこのような状態をあまり学習中に経験しておらず、学習が十分に進んでいないためと考えられる。しかし、限られた資源をより経験する状態に対して用いるという観点からは合理的であると考えられる。TD-based の強化学習の場合は、状態評価値の等高線の間隔が、手先が目標物に近いと狭く、遠いと広くなる。

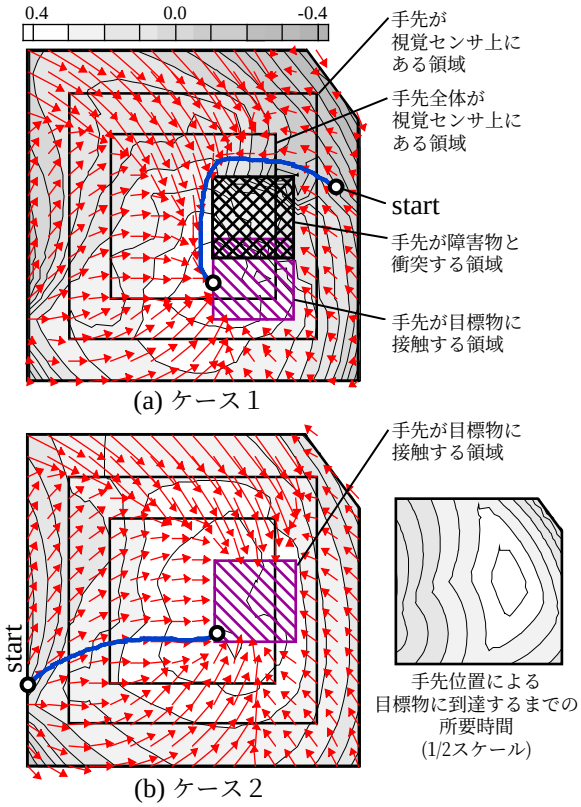


図 5 学習後の手先位置に対する状態評価値と手先の動作ベクトルおよびケース 2 の場合の最小所要時間

5.2 中間層表現の解析

ここでは、ニューラルネットの中間層が、目標物や手先や障害物の位置をそれぞれ独立して表現しているのか、また視覚と関節角の情報をどのように使っているかを観察する。まず、強化学習前と後のニューラルネットを用意し、**図 3** のように、それぞれ別の 2 つの出力ニューロンを付加する。中間層から出力層への重み値はすべて 0.0 とし、教師あり学習で付加したニューロンのみを学習する。つまり、入力層から中間層への結合の重み値のみが利用されることになる。目標物を視覚センサの 4 つの角に順番に置き、手先の位置は、視覚センサで見え、目標物と重ならない範囲でランダムに決定する。教師信号は、視覚センサ上の目標物の位置 x, y を正規化したものと、関節角θ_1, θ_2 の 2 つを用意した。そして、強化学習の適用前後でニューラルネットの出力を比較した。

図 6 に、視覚センサ上の位置を教師とした場合の x と出力 1 との関係と、関節角を教師信号とした場合の、θ_2 と出力 2 の関係を示した。出力は、100 個の目標物と手先の位置の組をランダムに選んでプロットした。いずれの場合も、強化学習前の方が x または θ_2 と出力が 1 対 1 の関係に近くなった。学習中には、目標物は視覚センサの四隅にしか現れていないため、強化学習前で

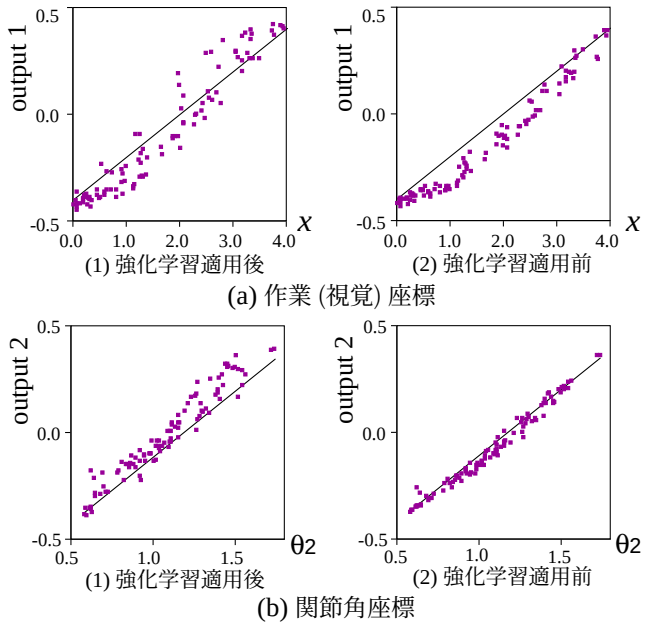


図 6 教師あり学習の結果から見た、強化学習前後での中間層における目標物位置の表現の比較

は、目標物が視覚センサの四隅以外に位置する場合に、その位置を表すような出力は得られないだろうと予測したが、実際には予測と大きく違った結果となった。学習の初期には、予測どおり、強化学習前のものは、センサの四隅以外で 0.0 に近い値であり、強化学習後では、**図 6** に近い出力となっていた。ところが、強化学習前のものは、学習が進むにつれ、学習中に目標物が現れない位置でも出力がその位置を表現するようになっていった。これは、画像から目標物の位置を知るためには、手先の画像の影響を取り除かなければならないが、その際に、関節角度から手先の画像への関係を学習し、視覚センサの各セルの位置関係が間接的に学習されたためと考えられる。また、手先と目標物が視覚センサ上で区別できないことも、その効果を大きくしたと考えられる。一方、強化学習後では、強化学習を通して手先と目標物の情報を統合した形で中間層で表現されているため、出力は、目標物の位置だけによらなかったと考えられる。

次に、手先位置の表現を調べた。視覚センサ上の手先の位置 x と y を正規化したものを教師信号として 2 つの出力を学習させた。ここでは、目標物と障害物は視覚センサ上に現れないものとし、手先の位置をランダムに決定した。そして、学習後に、**図 7** の 4 つの場合について、強化学習の適用前後で手先位置 x, y に対する出力を比較した。最初は、学習後の出力、2 番目は、視覚センサ上の手先位置を真ん中に固定し、関節角だけ変化させた場合、3 番目は、逆に、関節角を手先が視覚センサの真ん中に来るところで固定して視覚センサ上の位置だけ変化させた場合を示す。また、4 つ目は、手先位置を視覚センサ外まで広げて学習したものである。これらを見ると、強化学習の適用前は、x の表現には主に関節角が、yの表現には主に視覚センサ信号がかかわっていることがわかる。一方、強化学習後は、障害物が出現する場所の近くに手先がある場合を関節角が重点的に表現し、視覚センサ信号による表現は解釈が難しい。また、手先の存在範囲を視覚センサの外まで広げると、強化学習前では、主に関節角で手先位置を表現しているのに対し、強化学習後は、視覚センサ信号の影響が大きくなることがわかる。このことから、中間層の表現が、強化学習を通して適応的に変化し、視覚センサと関節角を統合したものとなっていることがわかる。手先が視覚センサ外に出ると、手先位置は関節角を使って表現することしかできない。よって、強化学習前では、関節角ですべての手先位置を表現するようになるが、リーチングの学習では、関節角と視覚の関係がわからないと目標物の位置がわからないため、手先の位置も両者を統合した形になると考えられる。ニューラルネットの初期値を変えてシミュレーションしても、同様の傾向がみられた。

6. 生体との関連性

Graziano らは、サルの腹側運動前野で見られる bimodal neuron、つまり、触覚と視覚の両方に反応するニューロンのうち、手先に触覚の受容野を持つもののほとんどは、視覚の受容野が、目を動かしても動かず、手を動かすと一緒に動くことを示している[9]。そして、これらがリーチング運動を行うために有効であろうと述べている。また、入来は、頭頂間溝と呼ばれる部位において、手が到達する範囲の視覚刺激に反応する多数のニューロンにおいて、サルが道具を持つことによって、視覚の受容野がその分増大し、道具使用を中止すると、たとえ道具を手に持っていても 1〜5 分後には受容野が小さくなったと報告している[10]。また、道具を使っても届かないところにエサが置かれると、サルは、試してみることなく、即座にエサを取ることをあきらめることから、サルは、視覚情報から到達できる距離かどうか判断できるとしている。本稿で述べたニューラルネットの状態評価値の出力は、リーチングまでの所要時間をコードしているため、**図 5** のように、目標物の位置が手先に近いほど大きな値となる。また、道具を持てばリーチングできる領域が広がることから上記のような受容野の変化も説明できる。つまり、上記のようなニューロンは、エサをとるというタスクを強化学習に基づいて学習することによって得られるものではないかと考えることができる。

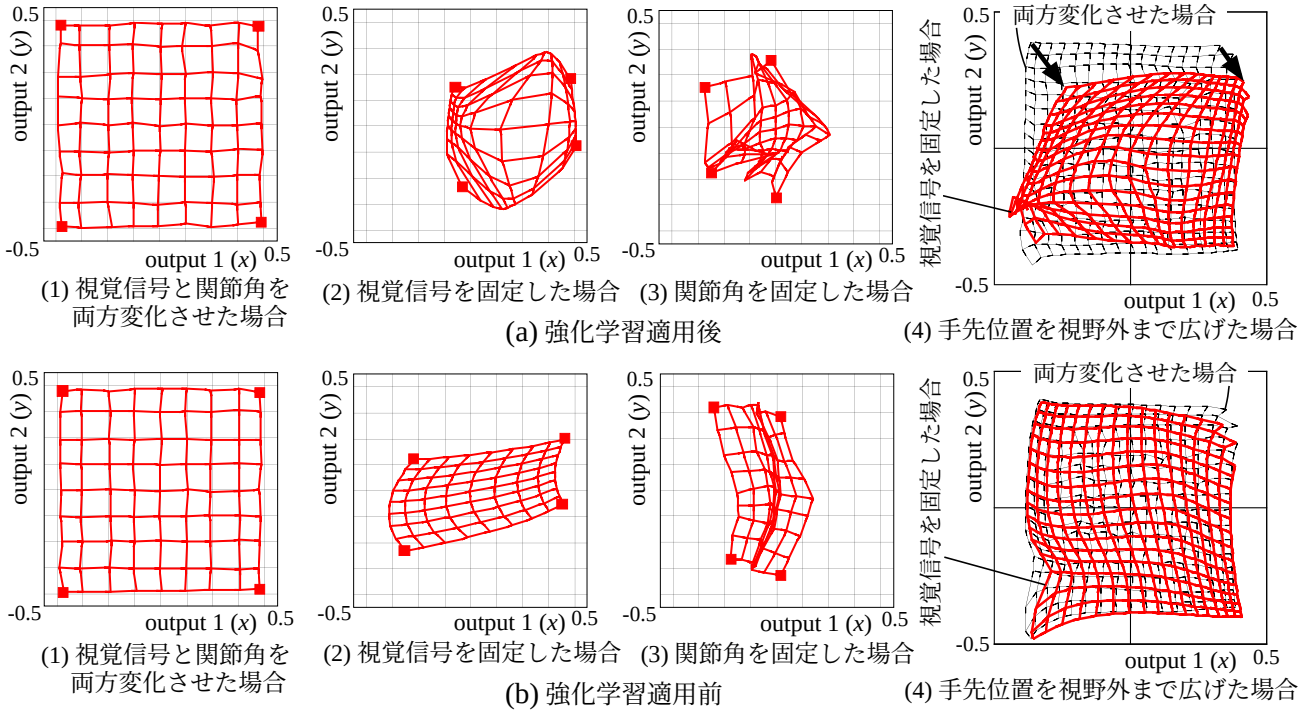


図 7 教師あり学習の結果から見た、強化学習前後での中間層における手先位置の表現の比較

7. 結論

強化学習とニューラルネットの組合わせにより、生の視覚センサ信号と関節角の信号を入力としてリーチングタスクの学習が行えることを示した。視覚センサ上の目標物、手先、障害物が区別がつかなくても学習を行うことができた。学習後の中間層の表現の解析から、目標物と手先の位置を別々に表現しているのではなく、統合した形で表現していること、さらに、目標物の位置を、視覚センサ信号、または関節角度だけで表現しているのではなく、両者を統合した形で表現していることがわかった。

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。

参考文献

1) Asada, M., et al.: Purposive Behavior Acquisition for a Real Robot by Vision-Based Reinforcement Learning, *Machine Learning*, **24**, 279–303 (1996).
2) Anderson, C. W.: Learning to Control an Inverted Pendulum Using Neural Networks, *IEEE Control System Magazine*, **9**, 31–37 (1989).
3) Shibata, K., Okabe, Y. & Ito, K: Direct-Vision-Based Reinforcement Learning in “Going to an Target” Task with an Obstacle and with a Variety of Target Sizes, *Proc. of Int’l Conf. on Neural Networks and Their Applications (NEURAP) ’98*, pp. 95–102 (1998).
4) 柴田克成, 岡部洋一, 伊藤宏司: ニューラルネットを用いた Direct-Vision-Based 強化学習, 第 4 回創発システムシンポジウム資料 (1998)
5) R. Held & J. A. Bauer: Visually Guided Reaching in Infant Monkeys after Restricted Rearing, *SIENCE*, **155**, pp. 718–720 (1967).
6) M. Jägersand & R. Nelson: Adaptive Differential Visual Feedback for Uncalibrated Hand-Eye Coordination and Motor Control, TR 579, Univ. of Rochester (1994).
7) Barto, A. G., Sutton, R. S. & Anderson, C. W.: Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems, *IEEE Trans. of SMC*, **13**, pp. 835–846 (1983).
8) 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, Vol. 117-C, No.9, pp. 1291–1299 (1997)
9) Graziano, M. S. A, Yap, G. S., & Gross, C. G.: Coding of Visual Space by Premotor Neurons, *Science*, **266**, pp. 1054–1057 (1994)
10) 入来篤史: サルの道具使用と身体像, 神経進歩, Vol. 42, No. 1, pp. 98–105 (1998)