H-001

複雑な運動パターン獲得のための強化学習モデル
A reinforcement learning for acquisition of complicated trajectory

徳永 憲市 †
Ken-ichi Tokunaga

和田 安弘 †
Yasuhiro Wada

1. はじめに

人の複雑な運動パターンのひとつである書字軌道を再現するモデルが，以前の研究で提案されている [1]．これらのモデルは運動パターンの獲得のために試行錯誤を繰り返すモデルではなく，与えられたアルゴリズムによって運動パターンを獲得するモデルであるため，人が複雑な運動パターンを獲得する際の学習過程を示しているとは考えにくい，という指摘がある．Grossberg & Paine[2] は，Wada & Kawato[1] の経由点学習アルゴリズムは，示唆的な学習アルゴリズムであると述べているが，一方，バイオロジカルに合理的でない点を指摘している．

本研究では，経由点表現を用いて，複雑な運動パターンを試行錯誤を繰り返しながら学習して獲得するモデルを提案する．最初に，我々は人の書字運動の獲得過程を行動実験によって確認し，その特徴を調査した．行動実験結果を基に構築したモデルは，運動パターンを再現するための適切な経由点位置を学習し，運動パターンを獲得するものである．本モデルは，強化学習のひとつである actor-critic 法 [3] によって，試行錯誤的に学習を行うモデルである．

2. 運動パターン獲得実験

2.1 実験方法

人が複雑な運動パターンを獲得する過程を計測し，その特徴を調査するため，以下の手順で行動実験を行った．

2.1.1 実験環境

実験環境を Fig.1 に示す．被験者は肩，手首を固定した状態で計測マーカのついたペンを手に持ち，机の前に座って実験を行った。Fig.1 左に示すように，計測マーカ (3 点) の位置を位置計測装置 (OPTOTRAK) によって計測し，計測したマーカの位置からペン先の位置を計算した．机の上に CRT を配置し，CRT 上には現在のペン先の位置および教示軌道を表示した．教示軌道の大きさは縦 250mm，横 130mm であり，Fig.2 の点線で示す形状であった．

2.1.2 実験手順

被験者には，CRT 上の教示軌道を可能な限り正確になぞることと，指示した目標運動時間で軌道を描くことを，実験開始前に指示した．実験中は 1 試行ごとに、運動時間および測定軌道を被験者にフィードバックした．目標運動時間は 5 秒とし，100 試行の測定を行った．実験中はサンプリング周波数 250Hz でデータを計測し，解析時にはカットオフ周波数 10Hz の 4 次のバターワースフィルタによってフィルタリングしたデータを用いた．

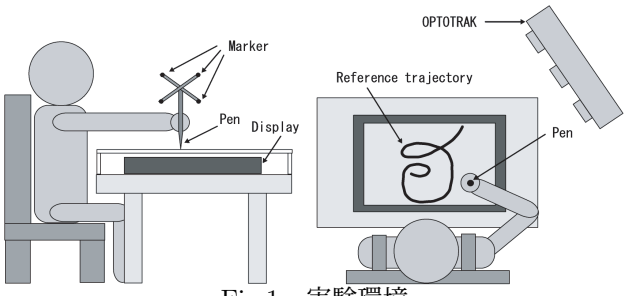

Fig.1　実験環境

2.2 実験結果

4 名の男性を被験者として実験を行った．Fig.2 に，1 名の被験者の実験開始時 (1 試行目) と実験終了時 (100 試行目) の測定軌道を示す．また Fig.3 に，実験開始時 (1 試行目) と実験終了時 (100 試行目) の接線速度波形を，運動時間および接線速度をそれぞれの最大値で正規化して示す．ここで，Fig.2 の測定軌道上に ◯ 印で示した点は，Fig.3 の速度波形における極小点の位置である．

Fig.2 および Fig.3 では，接線速度波形の極小点個数が学習によって減少していくことが示唆された．ここで，接線速度の極小点位置は，教示軌道を再現する上で重要な点であると考えられる [4]．従って，本研究において，極小点位置は教示軌道を再現する場合の経由点に相当すると仮定する．

次に，教示軌道と測定軌道間の位置誤差について，全被験者の平均推移を Fig.4，速度波形の極小点個数についての全被験者の平均推移を Fig.5 に示す．これらの結果から，位置誤差および接線速度の極小点数は，学習試行を重ねるに従って減少していくことが確認できた．

従って，人は複雑な運動パターンを学習する過程で，運動パターンを再現するために必要な経由点位置を学習し，さらにその個数が適切になるように学習していることを示唆していると考えられる．

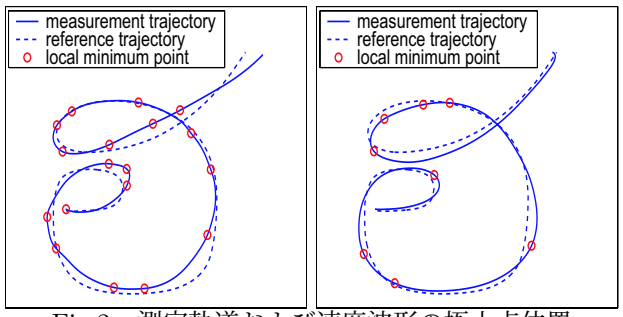

Fig.2　測定軌道および速度波形の極小点位置
(左:1 試行目，右:100 試行目)

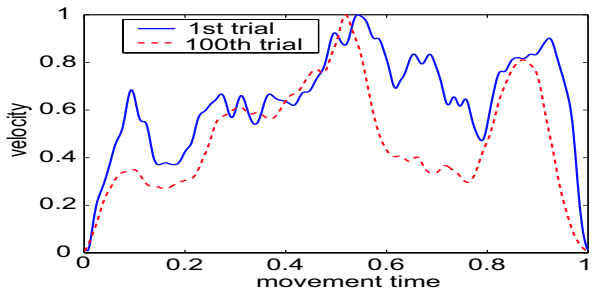

Fig.3　接線速度波形

†長岡技術科学大学，Nagaoka University of Technology

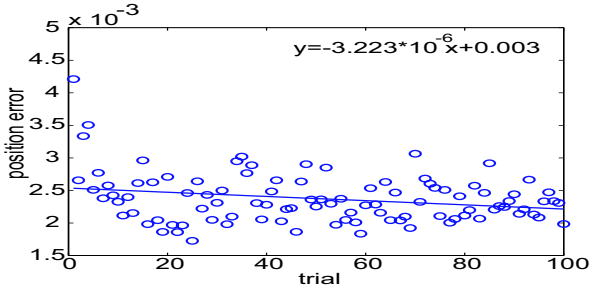

Fig.4　位置誤差の推移

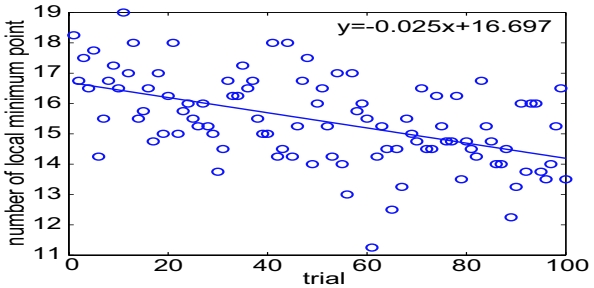

Fig.5　接線速度波形の極小点数の推移

3.　運動パターン獲得モデル

次に，前節の行動実験の結果を基にした運動パターン獲得モデルを提案する．学習モデルの構成を Fig.6 に示す．本モデルは，運動パターンの表現である経由点を選択するエージェント，経由点から実際の軌道を生成する軌道生成機構，生成した軌道の評価を行う評価器の 3 要素から構成されている．各要素の詳細は 3.2 節に示す．

3.1　運動パターン獲得方法

本モデルは，以下の手順で運動パターンを獲得する．

まず獲得の初期段階では，運動時間および距離の短いプリミティブ運動を繰り返し，それらの運動をつなぎ合わせることによって，教示された運動パターン全体を再現する．ここでプリミティブ運動とは，再現した運動パターンを構成する運動を指し，ここでは始点と終点間の 2 点間運動を指す．この段階においては，運動パターンを再現するために多くの経由点を必要とし，再現された運動パターンは滑らかではないが，教示運動パターン全体の形状を獲得することができる．教示運動パターンと再現した運動パターンとの間の位置誤差が閾値以下となった段階で，次の段階へと移る．

次の段階では，モデルが運動パターンをより滑らかに再現できるようにするため，1 度のプリミティブ運動で複数の経由点を通過する運動を行う．これによって，運動パターンの形状を再現するために必要なプリミティブ運動数や，運動パターン再現に必要な経由点数が減少する．モデルの学習経過に合わせて，1 度のプリミティブ運動に用いる経由点数を徐々に増加させていき，最終的には教示運動パターンを 1 度の運動で再現できるようになるまで経由点数を増加させる．このようにして獲得した運動パターンは，必要最小限の経由点のみによって再現される．

以上の手順で運動パターンを獲得することによって，行動実験の結果と同様に，学習過程で適切な経由点個数と経由点位置を獲得し，滑らかな運動パターンが再現できると考えられる．

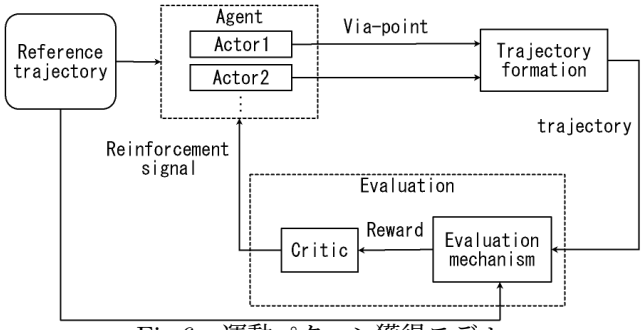

Fig.6　運動パターン獲得モデル

3.2　モデルの構成要素

3.2.1　エージェント

エージェントは，与えられた教示軌道を参照し，その軌道を再現するために必要な経由点を推定する．1 つの actor で 1 つの経由点を推定するため，actor は推定する点数分用意する．それぞれの actor は同じ構成をしており，AdaptiveGRBF [5] によって表現する．actor への入力を，l 回目のプリミティブ運動と，教示軌道上から取り出した経由点候補のインデックス k からなる 2 変数 (l,k) とすると，基底関数 $b_{i,j}(l,k)$ は以下のように定義される．

$$b^A_{i,j}(l,k) = \frac{a^A_{i,j}(l,k)}{\sum_{m=1}^{S}\sum_{n=1}^{K} a^A_{m,n}(l,k)} \tag{1}$$

$$a^A_{i,j}(l,k) = exp\left(-\frac{M^A_i(l-c^A_i)^2}{2} - \frac{M^A_j(k-c^A_j)^2}{2}\right) \tag{2}$$

ここで，S は教示軌道を再現するために繰り返したプリミティブ運動の総数，K は教示軌道のサンプル総数である．c^A_i, c^A_j はユニットの中心値，M^A_i, M^A_i はユニットの形状を決める定数である．このとき，l 回目のプリミティブ運動での経由点候補 k の選択確率 $Prob(l,k)$ を以下の式で計算する．

$$Prob(l,k) = \sum_{m=1}^{S}\sum_{n=1}^{K} w^A_{m,n} b^A_{m,n}(l,k) \tag{3}$$

ここで，w^A は基底関数の重みである．この確率分布を元に，運動の始点との距離を考慮して経由点を推定する．従って，w^A を学習することによって，モデルは教示軌道を再現できる経由点を獲得できるようになる．

3.2.2　軌道生成器

軌道生成器は，エージェントから得られた始点，経由点，終点の情報を元に軌道を生成する．本研究では，滑らかさの最適化規範である指令トルク変化最小規範 [6] に基づいて軌道を生成する．この規範は，各関節に生じるトルクの時間的変化を最小とする軌道を生成する運動規範である．その評価関数は以下のように表される．

$$C_{CTC} = \frac{1}{2}\int_0^{t_f}\sum_{i=1}^{N}\left(\frac{d\tau_i}{dt}\right)^2 dt \tag{4}$$

ここで t_f は運動時間，N は運動に関係する関節の数，τ_i は関節に発生する指令トルクを示している．本研究では，軌道生成器として FIRM[7] を用いる．

3.2.3 評価器

評価機構は軌道生成器によって得られた生成軌道と教示軌道との間の誤差に基づいて評価を行い，報酬 r_i を計算する．

$$r_i = exp\left\{-\alpha \int_0^{T_i} \sqrt{(x_i(t)-\tilde{x_i}(t))^2+(y_i(t)-\tilde{y_i}(t))^2}dt\right\} \quad (5)$$

ここで，T_i は i 回目のプリミティブ運動の運動時間，$(x_i(t), y_i(t))$ は生成軌道を表す．また，$(\tilde{x_i}(t), \tilde{y_i}(t))$ は教示軌道を表す．

この報酬 r_i は生成軌道に対する評価のみなので，i 回目のプリミティブ運動で得られた報酬 r_i と $i+1$ 回目以降の運動で得られた報酬 $r_{i+1}, r_{i+2}\cdots$ の累計 $V(i)$ を以下のように定義する．

$$\begin{aligned} V(i) &= r_i + \gamma r_{i+1} + \gamma^2 r_{i+2} + \cdots \\ &= r_i + \gamma V(i+1) \qquad (0 \le \gamma \le 1) \end{aligned} \quad (6)$$

ここで γ は割引率である．この $V(i)$ を critic を用いて予測する．critic も actor と同様に AdaptiveGRBF [5] によって表現される．critic への入力はプリミティブ運動の回数 i であり，出力 $V(i)$ は以下の式によって計算される．

$$V(i) = \textstyle\sum_{m=1}^{S} w_m^C b_k^C(i) \quad (7)$$

$$b_k^C(i) = \frac{a_k^C(i)}{\sum_{m=1}^{S} a_m^C(i)} \quad (8)$$

$$a_k^C(i) = exp\left\{-\frac{(M_k^C(i-c_k^C))^2}{2}\right\} \quad (9)$$

ここで，c_k^C はユニットの中心値，M_k^C はユニットの形状を決める行列，w^C は基底関数の重みである．このとき，評価の誤差 E_i を以下のように定義する．

$$E_i = r_i + \gamma V(i+1) - V(i) \quad (10)$$

E_i が正の場合は，actor が選択した経由点によってより高い評価が得られた事を意味し，負の場合はその逆を意味する．この E_i を強化信号として用いて actor の学習を行う．

3.3 モデルの更新方法

生成軌道の評価結果に基づいて経由点の選択確率を更新するために，actor 内の重み w^A を更新する．w^A は評価の誤差 (強化信号)E_i を用いて更新する．i 回目のプリミティブ運動で選択した点のインデックス j に対応する重み $w_{i,j}^A$ の更新量 $\triangle w_{i,j}^A$ を，以下のように計算する．

$$\triangle w_{i,j}^A = \eta_a \cdot E_i \cdot b_{i,j}^A(i,j) \quad (11)$$

ここで，η_a は学習係数である．ここで E_i が正の場合，つまり選択した点によって教示軌道をうまく再現できた場合，選択した点から離れた位置にある点については，選択確率を減少させるように学習を行うために，$j+1 \le k \le K$ の範囲の点 k に対応する重みの更新量 $\triangle w_{i,k}^A$ を以下のように計算する．

$$\triangle w_{i,k}^A = -\frac{G}{1+exp\{-((k-j)-p)/T)\}} \quad (12)$$

ここで，T は関数の変化の緩急を決めるパラメータ，p は変化する位置を決めるパラメータ，G は更新量を決める定数である．

一方 critic については，評価の誤差 E_i が 0 となるように，最急降下法を用いて重み w^C の学習を行う．

$$\frac{\partial(r_i+\gamma V(i+1)-V(i))^2}{\partial w_k^C} \propto \eta_c E_i b_k^C(i) = \triangle w_k^C \quad (13)$$

ここで，η_c は学習係数である．

4. シミュレーション

提案する学習モデルを用いて，実際に運動パターンを学習したシミュレーション結果を以下に示す．今回は，教示軌道としてアルファベットの ‘a’ を与えた．学習時に用いたパラメータを Table1 に示す．

Table1　学習に用いたパラメータ

K	130	N	2
M_i^A	10.0	M_j^A	12.0
γ	0.20	α	800
M_k^C	10.0	η_a	2.00
η_c	0.15	G	5.0
p	30	T	6.0

学習前半は，モデルは 2 点間運動を繰り返しながら軌道を再現する．Fig.7 に 2 点間到達運動の場合の教示軌道との位置誤差の推移を示す．学習開始直後から位置誤差は大きく減少し，約 100 回程度の学習試行で位置誤差は十分に小さくなる．Fig.8 に 2 点間到達運動を繰り返して学習する場合の学習開始時の軌道と 87 試行後の生成軌道を示す．ここで，経由点位置は生成軌道上に ◯ 印で示す．この段階においては，87 試行の学習後のモデルは，22 回のプリミティブ運動を繰り返して教示軌道を再現しているため，教示軌道との誤差は少なくなるが生成された軌道は滑らかではない．

位置誤差が十分に小さくなった段階で，軌道生成に複数の経由点を用いることによって，モデルがより滑らかな運動を獲得するように学習を行う．複数の経由点を通過して生成した軌道，およびに軌道生成に用いた経由点位置を Fig.9 に示す，複数の経由点を通過する運動によって軌道が生成されるため，生成軌道は 2 点間運動を繰り返す場合よりも滑らかな軌道となり，人の書字運動に似た軌道を獲得することが出来た．また，用いる経由点数を増加させることによってプリミティブ運動数は減少し，用いる経由点数を 10 点まで増加させた場合に，モデルは教示軌道の始点から終点までを 1 回の運動で再現できた．

学習全体の位置誤差の推移を Fig.10 に示す．ここで，各プロットは 5 試行分の平均を示している．モデルの学習によって，位置誤差は定性的に減少していくことが確認できた．また，モデルが Fig.9 右の軌道を獲得するまでに，約 300 試行程度の学習試行が必要であった．

次に，軌道全体を再現するために必要な経由点数の推移を Fig.11 に示す．ここで各プロットは，2 点間運動の場合，5 点の経由点を用いた場合，10 点の経由点を用いた場合の各学習区間で平均した値を示す．軌道全体を再現するために必要な経由点数は徐々に減少していくことが確認できた．

Fig.12 には学習初期と後期の接線速度波形を示す．ここで，運動時間と接線速度は正規化して示してある．学習初期においては，速度波形の極小点が多く存在していることが確認できる．これは学習初期において，モデルは 2 点間運動を繰り返すことによって教示軌道を再現しているからである．学習後半では，複数の経由点を用いて軌道を再現しているため，教示軌道の始点から終点までを 1 度の運動で再現できるようになる．従って，学習によって速度波形の極小点が減少してくる．

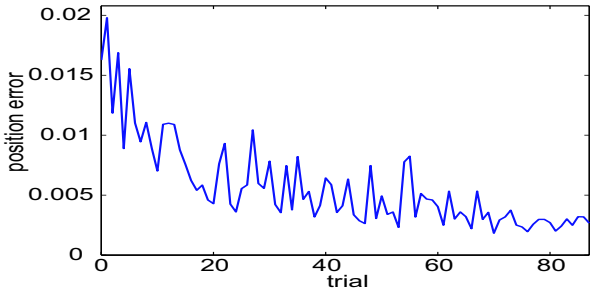

Fig.7　位置誤差の推移 (2 点間運動)

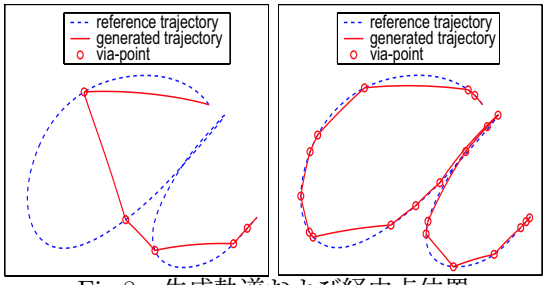

Fig.8　生成軌道および経由点位置
(左:1 試行目, 右:87 試行目)

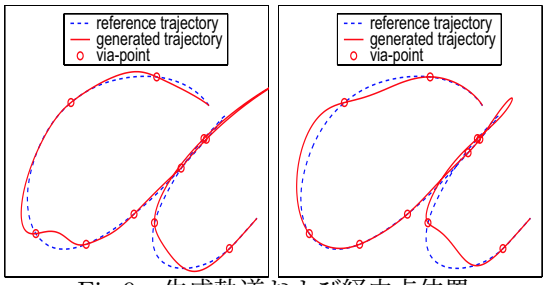

Fig.9　生成軌道および経由点位置
(左:経由点 5 点, 右:経由点 10 点)

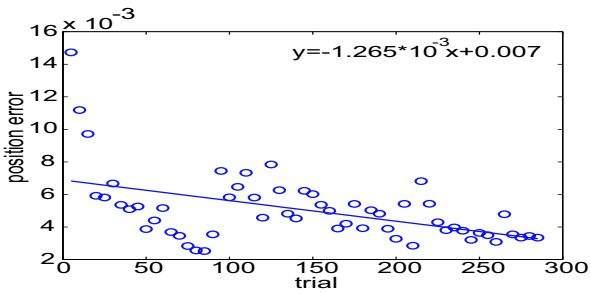

Fig.10　学習全体の位置誤差の推移

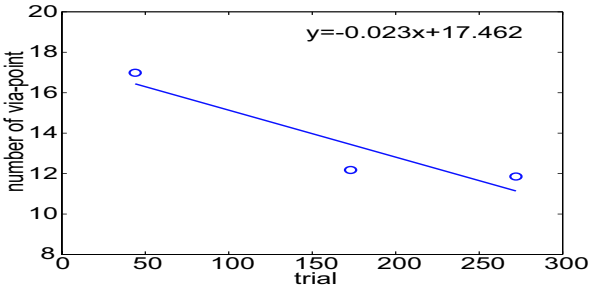

Fig.11　軌道再現に必要な経由点数の推移

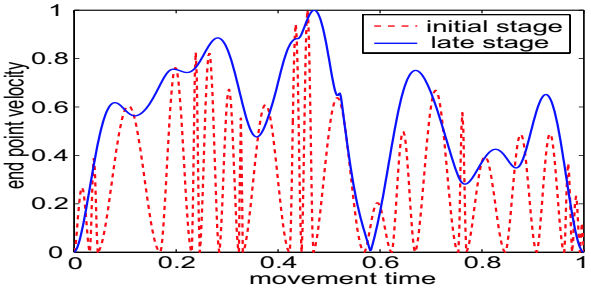

Fig.12　接線速度波形

5.　まとめ

本研究では，人の運動パターン獲得の過程を行動実験によって観察し，その結果を基にして，強化学習を用いた運動パターンの獲得モデルを提案した．本モデルによってシミュレーションを行った結果，学習によって複雑な運動パターンを再現可能な経由点情報を獲得できることが確認できた．このモデルの利点は，比較的少ない学習試行数で運動パターンを獲得できることである．また，モデルによるシミュレーション結果は，人による実験結果とほぼ一致しており，モデルの妥当性が示唆されていると考えられるが，今後，人の計測データとの詳細な比較検討が重要な課題である．

本モデルを応用する事によって，例えば人が手本となる運動パターンを与えた場合に，模倣によって運動パターンを学習するシステム等に利用することが可能と考えられる．

参考文献

[1] Wada, Y. and M. Kawato, A theory for cursive handwriting based on the minimization principle. Biological Cybernetics, 73(1):3-13.1995.

[2] Grossberg, S., Paine, R.W., A neural model of cortico-cerebellar interactions during attentive imitation and predictive learning of sequential handwriting movements. Neural Networks, 13(8-9),999-1046, 2000.

[3] Barto, A.G., Sutton, R.S., Anderson, C.W., Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Trans. Syst., Man, and Cybernetics, 13:834-846, 1983.

[4] 和田安弘，小池康晴，川人光男，連続書字運動の計算論的モデル，通学論，J76-D-Ⅱ，pp.2400-2410, (1993)

[5] Morimoto J., Doya K. Renforcement learning of dynamic motor sequence: Learning to stand up. Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems,3,1721-1726, 1998.

[6] Nakano, E., Imamizu, H., Osu, R., Uno, Y., Gomi, H., Yoshioka, T., Kawato, M., Quantitative examinations of internal representations for arm trajectory planning: minimum commanded torque change model. Journal of Neurophysiology, 81(5):2140-2155, 1999.

[7] Wada, Y., Kawato, M., A neural network model for arm trajectory formation using forward and inverse dynamics models. Neural Networks,6,919-932,1993.