大域結合カオス系に基づく強化学習による自律ロボットの行動獲得*

(第2報, 動的環境におけるナビゲーション行動の学習)

中村 陽一郎^{*1}, 大倉 和博^{*2}, 上田 完次^{*2}

Behavior Acquisition of an Autonomous Robot by Reinforcement Learning Based on Globally Coupled Chaotic System (2nd Report, Learning Navigational Behaviors in Dynamic Environment)

Yoichiro NAKAMURA, Kazuhiro OHKURA and Kanji UEDA

A framework for the acquisition of perception-based navigational behaviors by autonomous mobile robots is presented. A globally coupled chaotic system is applied for learning of reactive action-rules. The control architecture of a robot consists of chaotic elements each encoding an action-rule. The elements have the dynamics, which are designed so that the elements can collectively execute reinforcement learning. The proposed method has been tested in simulations dealing with a navigation task in changing environments. Simulation results indicate that the robot acquired behaviors such as light-seeking and wall-following. The robot became able to move along a wall with its head swinging reciprocally so as to keep perceiving the wall on its left side. This strategy is flexible enough to enable the robot to search for, keep following, and turn along the wall. The process of behavior acquisition is also discussed from the point of view of credit assignment.

Key Words: Mechatronics and Robotics, Moving Robot, Learning, Behavior Acquisition, Navigation, Reinforcement Learning, Chaos, Synchronization

1. はじめに

ナビゲーション・タスクを扱い, 自律移動ロボットに走行や衝突回避などの行動戦略を学習させることが試みられている。しかし, 動的環境, しかも連続的な空間でそれらの行動(behavior)を学習させた例は少ない。Venturini は, 学習の枠組みにクラシファイア・システムを用い, 複雑かつ変化する環境下において行動学習を行っている⁽¹⁾。ただし, 学習器には, 動作(action)に対する評価が, 目標との距離の変化に基づき1動作ごとに与えられており, 結果的に目標の位置をロボットに教示しておく必要がある。

本研究では, 強化学習法を用い, 環境について先験的知識を持たない自律移動ロボットに, より少ない評価でナビゲーションに関する行動戦略を獲得させる。前報では, 大域結合カオス系に基づく強化学習機構を自律移動ロボットに適用し, 静的環境下での行動学習を試みた⁽²⁾。本報では, 引き続き計算機実験を行い, エピソードごとに障害物の配置を変えることにより, 変化する環境下での行動学習を試みる。そして, 実際

に行動を導いている動作ルール列を調べ, ロボットが獲得した行動戦略について議論する。

2. 自律移動ロボットの制御アーキテクチャ

2・1 概要 ロボットの制御システムは, 観測した状態に対して即応的に動作を発生させることにより, 実際に経験した状態—動作のペアを動作ルールとして記憶していく。しかし, 即応的に発生した動作は, その場の状況と整合性があるとは限らない。そこで, 強化学習を行い, 状況と整合性のある動作を導くルールを獲得していく。

金子により大域結合写像として提案された大域結合カオス系は, 大域的な場を通じて相互作用するカオス要素群が, 集団的に同調して振動するクラスタリングや, カオス性により同調関係を動的に変化させるカオス遍歴などの性質を持つ⁽³⁾。これまでに, パラレル通信への応用が試みられているほか⁽⁴⁾, Steels は, 相互作用し合う複数の移動ロボットの挙動を, 大域結合写像として捉えている⁽⁵⁾。本研究では, 大域結合カオス系を行動の強化学習機構に用いている。具体的には, カオス要素の遍歴的挙動により多様な動作ルールをオンラインで発生し, 同調関係として記憶する。また, 同調関係どうしの相互作用を, 動作ルール間の競合に利用する。

* 原稿受付 1996年12月19日。

^{*1} 正員, 日立造船(株)技術研究所(〒551 大阪市大正区船町2-2-11), (神戸大学自然科学研究科)。

^{*2} 正員, 神戸大学工学部(〒657 神戸市灘区六甲台町1-1)。

Fig.1 Control architecture

2.2 状態の知覚と動作の規定 計算機実験に用いる自律移動ロボットは、2個の光センサと7個の近接センサ、および5個の衝突センサを持つ[図1]。制御システムには、光センサと近接センサの計9個の出力を2値化した9ビットのストリングが、状態入力として与えられる。

一方、ロボットの動作は、エフェクタの駆動操作 $O=\{o_1, o_2, \dots, o_{N_c}\}$ の関係付け(組み合わせ)により規定される。図1のロボットは、左・右それぞれの車輪の前方回転、両輪の後方回転の3種類の操作を持つ。操作の組み合わせは、次の7パターンになる：

$\{o_1\}, \{o_2\}, \{o_3\}, \{o_1, o_2\}, \{o_2, o_3\}, \{o_3, o_1\}, \{o_1, o_2, o_3\}$
 操作の組み合わせに応じて左右の車輪に回転角度差が生じると、前進、微速前進、後退、左・右回転、左・右旋回が差動機構によって得られる。

2.3 制御システムの構成 制御システムは、カオス要素の集合 E から構成される。 E は、カオス遍歴により互いに無限定な同調関係を形成できる要素(無限定要素)の集合 E_{IND} と、同調関係を限定された要素(限定要素)の集合 E_{DEF} からなる。

$$E = \{E_{IND}, E_{DEF}\} \dots \dots \dots (1)$$

カオス要素 $e_i \in E$ は、 D_i をダイナミクス、 X_i を内部状態変数、 mem_i を記憶値として次のように記述される。

$$e_i = \langle D_i, X_i, mem_i \rangle \dots \dots \dots (2)$$

状態入力に対して、要素間の同調を利用して各エフェクタ駆動操作 o_i の間に関係付けを行い、動作を決定する。 o_i には、カオス要素コラム C_i に対応させる。

Fig.2 Flow of reinforcement learning

$$C_i = \{e_{i1}, e_{i2}, \dots, e_{iN_i(s_i)} \mid e_{ij} \in E_{DEF}, e_{iN_i(s_i)} \in E_{IND}; i=1, \dots, N_c, j=1, \dots, N_i(s_i)-1\} \dots \dots \dots (3)$$

式(3)のように、各コラムには無限定要素が常時1個だけ存在する。初期状態では、各コラムには無限定要素1個しかない。状態 s_n におけるコラム C_i の要素数 $N_i(s_n)$ は、要素の再生と死滅により状態ごとに変わる。

状態 s_n において、カオス要素 $e_{il} \in C_i$ と $e_{jm} \in C_j$ の同調関係が、競合過程を経て唯一持続した場合、コラム C_i と C_j は、両者の間に関係が形成されたとして、対応する操作 o_i と o_j を同時に発火する。このように、要素間の同調関係は状態 s_n での動作ルールに相当する。

3. 大域結合カオス系による強化学習

図2に、強化学習の流れを示す。本文中の番号は、それぞれ図中の番号と対応している。

①学習開始時、各コラム内には無限定要素しか存在しない。センサ入力に対して、無限定要素は同調関係を変化させながら互いに競合する。②競合を経て同調関係が一つだけになった場合、その関係に対応した動作が発火する。③動作を発火させた無限定要素は、その状態でのセンサ入力ストリングを記憶するとともに、新たな状態—動作のルールとして同調関係をロックし、限定要素となる。④同時に、動作ルールの生成能力を維持するため、無限定要素をそれぞれのコラムに再生する。⑤すでに経験した状態では、その状態をストリングで記憶していた全同調関係が想起・活性化し、無限定要素の同調関係も含め、自己の同調だけを持続させるべく競合する。⑥競合に勝った同調関係が動作を発火する。⑦ロボットが目標状態に到達したときのみ報酬が与えられる。⑧報酬は、同調関係とそれに対応した動作の有効度として、要素群に分配されていく。⑨一定値以下に有効度を減らした要素は死滅する。

4. カオス要素のダイナミクス

以下では、式(4)のガウス型関数、式(5)のシグモイド関数、および式(6)のしきい値処理関数を用いる。

$$f(x, c, \theta) = \exp[-c(x - \theta)^2] \cdots \cdots \cdots (4)$$

$$g(x, c, \theta) = \frac{1}{1 + \exp[(\theta - x)/c]} \cdots \cdots \cdots (5)$$

$$h(x, \theta) = \begin{cases} 1: & \text{if } \theta < x \\ 0: & \text{otherwise} \end{cases} \cdots \cdots \cdots (6)$$

4.1 カオス要素の構成 式(2)で定義したカオス要素 $e_i \in E$ は、5 個の内部状態変数を有している。

$$X_i = \{x_i, u_i, b_i, r_i, d_i\} \cdots \cdots \cdots (7)$$

x_i : 振動子状態量 u_i : 同調関係の有効度

b_i : 振動出力荷重 r_i : 報酬分配率

d_i : 振動場 $\bar{x}$ との相関度

各変数を、ダイナミクス D_i , すなわち、式(9), (17), (18), (21), (22)の 5 個の離散時間方程式で更新する。

また、記憶値 mem_i は、次の 2 種類の値からなり、個々の同調関係を分離するために用いられる。

$$mem_i = \{Str_i, a_i\} \cdots \cdots \cdots (8)$$

Str_i : 状態ストリング

($Str_i = \{0, 1, \#\}^L$ # : Don't care symbol)

a_i : 動作識別コード ($a_i \in \{1, 2, \dots, 2^{N_c} - 1\}$)

L は、ストリング長を表す。限定要素の場合、 Str_i には再生された状態でのセンサ入力ストリングを与え、 a_i には要素 e_i の同調に対応して発火された動作の識別コードを与える。無限定要素では、 Str_i の全ビットに $\#$ を与え、 $a_i = 0$ とする。

4.2 基本ダイナミクス カオス要素の非線形振動子部分のダイナミクスを示す。

$$\left. \begin{aligned} x_i[t+1] &= c_0 f(x_i[t], c_1, \theta_x) - c_2 y_i[t] + c_3 x_i[t] \\ &\quad + \varepsilon(\bar{x} - x_i[t]) + A \sin(\bar{x} - mem_i') t + I_i S_i \cdots \cdots (9) \\ y_i[t+1] &= g(x_i[t+1], c_4, \theta_y) \end{aligned} \right\}$$

y_i は、 x_i のピークでパルスを出力する。式(9)第 1 式の第 4, 5, 6 項を、それぞれ結合項、想起入力項、活性制御項と呼ぶ。活性制御項の値によって、安定不動点の有無、つまり振動状態の活性・不活性が決まる。活性制御項の二つの一時変数は、次式で与えられる。

$$I_i = \begin{cases} 1: & \theta_m < m_i \\ 0: & \text{otherwise} \end{cases} \cdots \cdots \cdots (10)$$

$$S_i = g(I_i \hat{b}_i, c_5, \theta_b) \cdots \cdots \cdots (11)$$

式(10)の m_i は、状態ストリング $Str_i = \{str_i^1, \dots, str_i^L\}$ と、センサ入力ストリング $Sen = \{sen_1, \dots, sen_L\}$ とのマッチ率で、次式で計算する。 $\oplus$ は排他的論理和を表す。

$$m_i = \sum_{j=1}^L str_i^j \oplus \overline{sen_j} / L \cdots \cdots \cdots (12)$$

式(11)の $\hat{b}_i$ は、4.4 節で示す競合パラメータである。し

たがって、一定数以上のマッチ率をとり、競合中の要素は活性状態を維持して振動を持続する。新たに動作を発生させようとする無限定要素は、状態ストリングの全ビットが $\#$ なので、どの状態でも活性化する。

4.3 同調関係の形成と記憶 要素間に直接的な結合はなく、二つの大域場を介して相互作用がなされる。結合項は、次式で表される大域振動場 $\bar{x}$ との結合量を与える。 $\bar{x}$ は、系全体での荷重平均振動量である。

$$\bar{x} = \frac{\sum_{e_j \in E} b_j[t] x_j[t] / n_j}{\sum_{e_j \in E} b_j[t] / n_j} \cdots \cdots \cdots (13)$$

n_i は、要素 e_i が形成している同調関係の構成要素数である。各状態において活性化したカオス要素は、大域振動場を通じて全要素と相互作用する。無限定要素の場合、大域結合係数 ε は 0.08 であり、同調関係はカオス遍歴により動的に変化する。動作ルールとして同調関係を記憶する場合には、 $\varepsilon = 0.25$ として同調をロックする。このとき、以下のように想起入力項に変動を与え、異なる動作ルールどうしが同調するのを防ぐ。

同調して動作を発火した要素群は、互いに同一の状態ストリング Str_i と、動作識別コード a_i を記憶している。そこで、 Str_i を 10 進数にデコードしたうえで a_i と加算し、式(9)第 1 式の想起入力項に作用させる。

$$mem_i' = Str_i' + 0.1 a_i \cdots \cdots \cdots (14)$$

全要素が共通して影響を受ける大域振動場 $\bar{x}$ と、同調関係ごとに異なる mem_i' の差が、想起入力項の正弦関数に入力される。ゆえに、等しい記憶値 $mem_i = \{Str_i, a_i\}$ を持つ要素群は、想起入力項に等しい変動を受けるため、同調を維持できる。他方、記憶値が異なれば同調アトラクタ間に偏差が生じる。これはカオス性によって拡大されるため、同調関係どうしは分離される。

4.4 同調関係の競合 活性化した要素群の間には、振動出力荷重 b_i の平均場[式(15)]を通じた相互作用により、同調関係の存続をかけた競合が生じる。

$$\bar{b}[t] = \frac{\sum_{e_j \in E} b_j[t] / n_j}{\sum_{e_j \in E} S_j[t] / n_j} \cdots \cdots \cdots (15)$$

この値と b_i の比として、競合パラメータ $\hat{b}_i$ を定める。

$$\hat{b}_i = \frac{b_i[t]}{\bar{b}[t]} \cdots \cdots \cdots (16)$$

競合パラメータは、式(11)で与えられる活性制御項に作用し、要素の活性・不活性を決める。 b_i および振動場との相関度 d_i は、それぞれ以下のように更新する。

$$b_i[t+1] = \frac{u_0 + \alpha \cdot m_i u_i[t]}{2} \cdot [\hat{b}_i(1 - \beta + y_i[t+1] d_i[t]) + \beta \cdot d_i[t]] I_i S_i \cdots \cdots (17)$$

$$d_i[t+1] = f(\bar{x} - x_i[t], c_6, 0) \lambda^{1 - \hat{b}_i^2} \cdot d_i[t](1 - y_i[t]) + y_i[t] \cdots \cdots (18)$$

式(17)から、有効度 u_i 、マッチ率 m_i 、および d_i が大きければ振動出力荷重 b_i も大きくなり、その要素の振動

Fig.3 Reproduction of generalized action-rules

の大域振動場支配率は式(13)から高まる。それに伴い式(18)から d_i が増加し、競合パラメータ $\hat{b}_i$ は一層増す。 $\hat{b}_i$ を減らした要素が不活性化する一方、競合の勝者だけが振動を持続し、同調している勝者の振動がそのまま大域振動場の振動となる。勝者要素は、この状態を次式で個別に評価し、同期した振動のピークで $F_i=1$ をとった瞬間、各コラムのエフェクタ操作を発火する。

$$F_i = h(\lambda^{[\bar{b}_i(t+1)-2\bar{b}_i(t)]}, \theta_f) \dots \dots \dots (19)$$

競合中の同調関係は、カオス性により周期を変えながら互いに引き込みあっている。有効度の小さい同調関係でも、短期的にせよ複数の位相を合わせて振動すれば大域振動場の支配率が上がるので、より有効度の大きい同調関係に競合で勝つ余地が生まれる。

4.5 汎化規則の発生 図3に、汎化性を持つ動作規則の再生過程を示す。センサ入力ストリングに状態ストリングが完全には一致していなくても ($\theta_m < m_i < 1.0$)、有効度が大きい限定要素は、競合に勝ち動作を発火する場合がある。このとき、その要素 e_i は親となり、新たに汎化要素を再生する。親要素は、自己の状態ストリングのうち、センサ入力ストリングにマッチしていないビットを#に置き換え、別の状態ストリング $Str_{Rep} = \{str_R^1, \dots, str_R^L\}$ を作る。

$$\text{for } j \in \{1, \dots, L\} \\ str_R^j = \begin{cases} \# & : str_i^j \neq sen_j \\ str_i^j & : str_i^j = sen_j \end{cases} \dots \dots \dots (20)$$

汎化要素は、ストリング Str_{Rep} 、および親と同じ動作識別コードを記憶する。ゆえに、汎化要素が形成する同調関係、すなわち汎化規則は、再生された状態と親要素が記憶している状態を含む、複数の状態においてマッチ率 1.0 で活性化できる。汎化規則の有効度

は、親の有効度 u_i とマッチ率 m_i から $u_i m_i$ とする。

4.6 信頼度割り当て 要素は、同調として記憶している動作規則の有効度を、次式で更新する。

$$u_i[t+1] = \{1 - c_f F_i h(r_i[t], \theta_f)\} u_i[t] + \Delta u + r_i[t] P \dots \dots \dots (21)$$

$$r_i[t+1] = \gamma^F r_i[t] + S_i F_i (1 - \gamma \cdot r_i[t]) \dots \dots \dots (22)$$

ペイオフ P には、有効度を増す報酬と有効度を減らす罰がある。ペイオフは、発火した動作列に対する評価として、式(21)にしたがい過去複数の状態での競合の勝者に分配率 r_i で分配される。この分配率は、状態遷移ごとに割引率 γ で削減される[式(22)]。また、デッドロックやループを避けるため、限定要素は動作発火時に発火コスト $c_f u_i$ を支払う。目標状態に到達した際にも消散率 η で有効度を削減するため、報酬の獲得に寄与しない要素は徐々に有効度を減らし、しきい値以下 ($u_i < u_{min}$) になった要素は死滅する。一方、動作ルール間の有効度伝播量 Δu は、前報で比較した3方式のうち、最も良好な結果を得た次式で定める。

$$\begin{cases} \Delta u = \delta u \cdot F_i \cdot h(r_i[t], \theta_f) & : 0 < \delta u \\ \Delta u = 0 & : \text{otherwise} \end{cases} \dots \dots \dots (23)$$

$$\delta u = (\bar{b}[t] - u_0) - \alpha \cdot u_i[t]$$

発火コストを課していることを考慮し、状態 s_{n-1} での勝者ルール $R(s_{n-1})$ の構成要素 $e_i \in R(s_{n-1})$ は、競合に用いた有効度 $\alpha \cdot u_i$ を減らさない。また、状態 s_n での勝者 $e_j \in R(s_n)$ からは、 $\alpha \cdot u_j < \alpha \cdot u_i (= \bar{b} - u_0)$ の場合にのみ、伝播量 $\Delta u (= \delta u)$ を受ける。

5. 計算機実験

ナビゲーション・タスクを対象に、第2章で示した自律移動ロボットを用いて計算機実験を行う。

5.1 タスクの記述 試行では、スタートからゴールである光源到達までを1エピソードとし、ロボットが光源に到達した場合と、500回動作しても光源に到達しない場合に、ロボットをスタート地点に戻してエピソードを更新する。更新の際には、図4および図8に示したような5種類の環境を、確率的に選択する。ある環境で獲得されつつある行動が、別の環境においても有効であるとは限らないので、行動が不安定化しやすく、環境が静的な場合よりも学習は難しくなる。

スタート地点は各エピソードで一定であるが、ロボットの方向を毎回ランダムにして置く。近接センサと光センサは、ロボット本体直径のそれぞれ1.25倍と2.25倍の有効レンジを持つ。したがって、ロボットは図4の円内で光源を感知できる。ただし、ロボットは光源をゴールとは認識していないので、感知した光源に向かうとは限らない。そこで、光源到達時には報酬を、衝突時には罰をそれぞれ与えて、行動を学習させ

Fig.4 Task space (at Episode 2)

Fig.5 Number of collisions

Fig.6 Number of the newly generated rules and the chaotic elements

る。これらのペイオフは、カオス要素に対して、報酬:40.0, 罰:有効度の 5%を基準に、各要素の分配率 r_i に応じて分配される。消散率 $\eta=0.96$, 発火コストは有効度の 1% ($c_f=0.01$)である。ダイナミクスの各定数は、有効度伝播率を $\alpha=0.2$ とした以外、前報と同一である。カオス要素は、100 個まで増殖可能とした。この限度を越えて再生が行われる場合、有効度が最も低い要素が置き換えられる。

5.2 実験結果 図4は、試行のエピソード2におけるロボットの移動軌跡を表している。図5は各エピソードにおける衝突回数、図6はカオス要素数と新しく発生された動作ルール数の変化をそれぞれ表している。また、前進、左旋回、右旋回それぞれの動作を導く全ルールのうち、支配的になった代表的なルール

(a) Move forward

(b) Turn left

(c) Turn right

Fig.7 Reproduction process and utility transition of the principal action-rules

集合について、動作ルールの発生状況と有効度の変化を図7に示す。試行の後半では、衝突回数も少なくななり、定常的に光源に到達して報酬を得られるようになった[図5]。それにつれて特定の動作ルールが発火するようになり、各エピソードにおける新規動作ルールの発生がほとんどなくなっている[図6]。発火する特定のカオス要素群に報酬が分配される一方、それ以外のカオス要素は、ゴール到達時に有効度を消散率 η で減少させるので死滅していき、カオス要素数はエピソード100において54個まで減少した。

図8には、いくつかのエピソードにおけるロボットの移動軌跡を示した。ロボットは、環境が変わっても、

Fig.8 Behavior of the robot

それに適応して光源に到達するような行動戦略を獲得していることがわかる。ロボットは、スタート地点への配置角がランダムであるにもかかわらず、必ず本体左側に壁を感知した状態を保つ形で移動するようになった。特に、左右の旋回を交互に繰り返しながら左壁沿いに移動している点が興味深い。そこで、学習の過程と獲得された行動を詳しく調べる。

この試行の前半では、壁を両側に感知した状態で前進動作を導くルール集合が有効度を高め、壁沿いに直進する行動が支配的になりかけた。しかし、壁沿いに直進しても、続いてコーナー部でうまく曲がれなければ、光源には近づけない。また、環境が図 8(b)の状況に変化したエピソードでは、領域 C-D-E-F においてロボットがデッドロック状態に陥り、支配的になりかけた行動を失う例がみられた。結局、壁沿い直進行動も安定せず、その後のエピソード 42 では、各状態に対

して左旋回を導く動作ルール集合が非常に強くなり、ロボットはループ的挙動を示した。動作ルールには発火コストを課しているため、このループ的挙動は解消されたが、続いてエピソード 44 では、逆に右旋回動作が支配的なループ的挙動が現れた[図 8(a)]。

図 7(a)からは、前進動作のルール集合が短期間で有効度を減らし、定着しなかったことがわかる。図 7(b)(c)をみると、エピソード 42 と 44 において、高い有効度を持つ親ルールから、それぞれ左旋回と右旋回の動作を導く数多くの汎化ルールが再生されており、ループ的挙動を招いたことがわかる。その後、左右の旋回動作を導く動作ルール集合の中から、最終エピソードまで存続する動作ルールが現れ、ロボットは両旋回動作を交互に繰り返す形で移動するようになった。

図 8(b)(c)(e)に関しては、一部の動作列について、観測された状態と、それに対して実際に発火した動作ル

Fig.9 Process of behavior acquisition

ールを示している。図 8(f)に表したセンサ配置および状態ストリングの構成と比べると、左壁沿いの移動に関して、ロボットはおおむね次のような行動戦略を獲得していることがわかる。(1)本体左側の近接センサ p_6 , p_7 がアクティブな状況では右旋回し、(2) p_7 だけがアクティブな状況では左旋回する。これらの戦略によって、以下の行動が実現されている。

- ・壁を探して左側に感知する状況を保つ行動
- ・壁沿いに曲がる形で衝突を回避する行動
- ・左右の旋回を交互に繰り返して前進する行動

図 8(b)の環境でも、左側の壁に沿って移動すれば、ロボットは領域 C-D-E-F に入らずに光源に接近できる。

5.3 考察 以上のように、本手法では、行動ベース型ロボット⁽⁶⁾の基本をなす、要素行動の創発が可能である。そこで、その獲得過程について議論する。

図 9 に表したように、試行では、はじめに光源接近行動が獲得される。次に、光源接近行動を構成している動作ルール(例えば、000000011→前進)からの有効度伝播、または直接的な報酬分配によって、衝突を回避する旋回や回転動作など、他の動作ルールが強化される。図 8 のように、左右の旋回動作のルールが交互に発火する場合、ロボットはうまく直進するうえに、式 (23)は、有効度が高い方のルールに他方が有効度を近づけるようにはたらくので、両旋回動作ルールを記憶している限られたカオス要素間で、有効度が相補的に伝播する。図 7(b)(c)において、左旋回ルール(2)(3)(4)と右旋回ルール(7)(11)それぞれの有効度が、比較的高い相関を保ちながら推移していることから、この両ルール集合間の協調的関係の存在が推察できる。このように、次々と生成されるカオス要素の中から、より高い有効度を持つ限られた要素だけが存続する結果、壁沿い移動に関する柔軟な行動戦略が創発されたといえる。壁沿い移動行動が定着し、報酬分配を受ける動作ルールも限られてくれば、スタート地点へのロボッ

トの配置角や環境の違いにより動作数や経路は変わっても、壁沿い移動→光源接近という行動順序は同じになる。こうして、ロボットは安定して光源に到達するようになる。

状態を離散的に捉えてリアクティブに動作を決めることに起因する状態の誤認、いわゆる Hidden state problem⁽⁷⁾への対処が課題である。そこで、獲得した行動をベースに、この問題に対処し得るような強化学習の枠組みを構築したい。

6. おわりに

大域結合カオス系に基づく強化学習機構を用い、自律移動ロボットに変化する環境下において行動学習させることを試みた。ナビゲーションを扱った計算機実験では、カオス系内に報酬獲得に寄与する要素が自己組織していく結果、ロボットがゴールへの到達に有効な行動を学習することを確認した。そして、壁沿い移動行動を導く動作ルール集合を調べ、左右の旋回を繰り返して前進する行動や、衝突を回避する行動が、左側に壁を感知する状況を保つ形でなされていることを示した。変化する環境に適応した行動を導く、リアクティブな動作ルールの抽出が、本手法により可能であると結論づけられる。

謝 辞

本研究の一部は、日本学術振興会未来開拓学術プロジェクト「創発的シンセシスの方法論」(JSPS-RFTF96P00702)、及び、文部省科学研究費重点領域研究「創発的機能形成のシステム理論」(264)の研究助成を受けてなされたことを記し、謝意を表します。

文 献

- (1) G. Venturini, Adaptation in dynamic environments through a minimal probability of exploration, *From Animals to Animals: Proc. of the 4th Inter. Conf. on Simulation of Adaptive Behavior*, (1996), 371-379.
- (2) 中村・大倉・上田, 大域結合カオス系に基づく強化学習による自律ロボットの行動獲得(第1報), 日本機械学会論文集 (1997年11月号掲載予定).
- (3) Kaneko K., Clustering, coding, switching, hierarchical ordering, and control in a network of chaotic elements, *Physica D*, **41**(1990), 137-172.
- (4) 桶谷・潮, 大域結合写像を用いたカオス通信法, 電子情報通信学会論文誌, **J79-A**, 8(1996), 1427-1432.
- (5) L. Steels, Intelligence-Dynamics and Representations, L. Steels ed. *The Biology and Technology of Intelligent Autonomous Agents*, (1995), Springer-Verlag.
- (6) R. A. Brooks, A Robust Layered Control System for a Mobile Robot, *IEEE Robotics and Automation*, **2-1**(1986), 14-23.
- (7) R. A. McCallum, Hidden State and Reinforcement Learning with Instance-Based State Identification, *IEEE Trans. on Systems, Man, and Cybernetics*, **B26-3**(1996), 396-407.