自律分散遊泳ロボットの障害物回避行動の獲得

飯 島 大 典* 俞 文 偉* 横 井 浩 史* 嘉 数 侑 昇*

Behavioral Acquisition of Obstacle Avoidance for a Distributed Autonomous Swimming Robot

Daisuke Iijima*, Wenwei Yu*, Hiroshi Yokoi* and Yukinori Kakazu*

This paper describes a design method of a robot system which consists on mechanically linked plural agents, so as to realize autonomous robots which can adaptively behave in the real world. A difficulty in designing distributed autonomous systems, is how to embed a dynamics for self-organizing environment-oriented-action-rules in the systems. So, we propose a robot system which controlled by a decision making method that has oscillator and learning method based on same object functions, and which has a structure of mechanically linked agents that are identically designed, in order that each agent actively interacts with other agents and the outer world. For verifying an usefulness of the system, behavioral acquiring tests in target approaching and obstacle avoidance were implemented by using Distributed Autonomous Swimming Robot that was constructed by the proposed way. Moreover, for learning efficiency in complex tasks such as obstacle avoidance, we proposed Switching Q-learning in which previously obtained action rules were effectively used. As a result, the robot acquired simple obstacle avoidance behavior, and time-sequential connections between acquired action rules were observed by interaction-based learning in a distributed autonomous system. That is, it was verified that the proposed system design method was one of the solutions of adaptive systems to complex environments.

Key Words: Distributed Autonomous System, Autonomous Robot, Adaptive Behavior, Learning, Obstacle Avoidance

1. は じ め に

近年、システムに求められる作業条件の複雑化に伴う技術的な限界から、自律分散システムが注目を集めている [1]。これは全体を統合する管理機構を持たず、構成要素が自律的に行動しながら、協調・競合的に相互作用しあい、全体として目的作業を達成するシステムのことである。また、その特徴としては、構成要素同士の相互作用のうちに自己組織的に動作ルールが動的に生成される可能性から、限られた作業環境にとどまらない複雑な環境への適応性や、システムの一部の故障時における動作ルールの再構成による目的作業の遂行性などが期待されている。

こうした有効性を踏まえて、自律ロボットの研究分野でも、自律分散型のモジュールロボット群を機械的に結合した構成のロボットを扱った研究がいくつか行われている [2]~[4]。これらの研究の中ではシステムの形態形成や再構成における自己組織能力や多様な形態での移動能力などが議論されているが、実際の動的環境下での行動設計に関する研究は行われていない。

このような機械的結合型の自律分散ロボットシステムの最大の利点の一つは、複数の可動部位がおのおのの意思決定機構により独立に制御されることによる、多様な動作パターンの生成可能性である。そのため、本研究では実環境で適応的に行動可能なロボットの実現を目指して、機械的に結合された同一構成のモジュールロボット（ここではこれをエージェントとする。エージェント；環境知覚，意思決定，行動が可能な独立した制御機構）群からなるロボットを対象とした新たな自律ロボット設計論の確立を試みている [5]~[7]。

これまでに我々は、提案する設計概念に基づいて構成された自律分散遊泳ロボットによる障害物のない状況での目標物への接近行動の獲得可能性 [5][6] や、一部のエージェントが故障した場合における目的行動の遂行可能性などを確認している [7]。そのため本論文では、まず提案する設計概念をさらに具体的なものにするために、いくつかの項目について議論を行う。また、自律ロボットにとって必然的な問題として、障害物回避問題を扱う。その方法としては、学習効率を考慮し、いったん障害物のない環境で得られた動作パターンをうまく利用し、付加的に新たな動作パターンが得られるように、動作ルールを切り替える方法を提案する。得られた動作パターンについては解析を行

原稿受付 2000 年 1 月 13 日

*北海道大学大学院工学研究科

*Graduate School of Engineering, Hokkaido University

い、分散制御システムにおける行動学習のメカニズムへの一考察を加える。

以降、2章ではエージェントの設計論について議論を行い、実環境で適応的に行動可能なシステムを構築するための設計概念と意思決定手法である適応振動子法について述べる。実験のための各種設定については、3章にて提案した手法の一つの具体化である自律分散遊泳ロボットをはじめ、導入したQ学習などの詳細を記す。実験内容については4章にて説明する。まず第一段階として、以前に行った障害物のない環境での目標物接近の行動獲得について触れ、続いて障害物回避のために動作ルールを切り替える学習法の提案とその有効性を議論したのち、行動獲得実験について述べる。それから5章においては特に分散制御システムにおける行動学習のメカニズムに焦点を絞り、提案したシステム設計法の妥当性や障害物回避の行動獲得の解析を試みる。最後に6章において、本論文の総括を行うものとする。

2. エージェント設計

2.1 適応行動学習に必要な設計条件

自律分散システムの設計における難点は、動作ルールが自己組織的に作られる仕組みをいかに埋め込むか、という点である。以降では、適応的な動作パターンが自律的に獲得されるために、システムの構成要素であるエージェントの設計について、対象とする実環境や機械的に結合された自律分散構造の特徴を踏まえて議論を行う。

2.1.1 実環境下での行動実現

実環境の特徴は、時間変動性を持ち、ノイズが多く、それゆえに不確実で不安定といった性質を有することである。これらの性質による影響は水中で行動する場合には特に顕著に現れる。例えばFig. 1のような水中の魚の行動は、そのときの魚のひれの動きだけでなく、本来の水の流れに加えて、それまでの魚の運動による慣性力や、周囲のあらゆる動作主体により発生された流れなどがすべてかかわった上で決定される。すなわちこのことより、実環境下での適応的な行動実現のためには、外界と物理的な（理想的には力学的な）相互作用を行いながら動作パターンを獲得してゆく仕組みが必要と考えられる。

2.1.2 機械的結合型の自律分散システムの構造利用

自律分散システムにおいて各エージェントが機械的に結合されている場合、個々のエージェントの動作は他のエージェントの動作に大きく影響を受ける。そのため、仮に目的関数が個別に設定された場合、どのエージェントにとっても、目的関数を満たすことは非常に困難となる。一方、目的関数を同一に設定した場合、いずれかのエージェントの受ける評価がすべてのエージェントに共通するため、協調的に振舞うことが予想され、効率的な動作獲得が期待される（Fig. 2）。

さらに、エージェントごとに独立した制御性を有する自律分散構造の利点の一つは、容易な規模の拡張性であるため、この利点を保持するには、エージェントの数や結合形態の変化に対応可能なように、個々のエージェントを、あらかじめ順序関係のない、機構および機能において均一なものとして設計することが必要である（Fig. 3）。

Fig. 1 Action in a complex environment

Fig. 2 Unifying object functions

Fig. 3 Countermeasure to various scales and structures of agents

2.2 提案するエージェント設計の概念

上記に議論した内容に基づき、本研究の提案するロボットシステムの設計概念を以下にまとめる。

- エージェントにはあらかじめ歩様や泳様といった特別な動作パターンは設定せず、外界との物理的な相互作用に基づく行動学習により獲得するものとする（このための意思決定手法を2.3節にて適応振動子法として提案する）。
- 機械的に結合されたシステムの行動学習の効率化のために、各エージェントには同一の目的関数を設定する。
- システムの規模や構造の多様性のために、各エージェントは順位的に優劣のない、ハードウェア的にもソフトウェア的にも同一の設計とする。

2.3 意思決定手法～適応振動子法～

複数エージェントにより構成されるロボットのための、外界との相互作用に基づく意思決定手法として、適応振動子法を提案しており[5]～[7]、一般式を式(1)に示す。左辺の第一項は、出力解の高階の微分値に重み和をとることで過去の長期的な出力解の影響を導入する微分演算子項、第二項は外界からエージェントへの様々な反作用要因（センサ情報や近傍エージェント情報、外部エネルギーなど）を受容する外界因子入力項、右辺はエージェントの内部エネルギーである。この式では、エージェントの出力の結果生じる状態変化を環境からの反作用要因

として導入し、それらと過去の長期的な出力の影響を考慮に入れることで、再び環境に対し作用を与える出力を決定する、という相互作用的なはたらきを期待している。さらに外部のエネルギーを取り入れて内部エネルギーとのバランスを保たせようとすることで、各パラメータ学習を促すことを特徴としている（実際には本論文では問題の簡単化のため、エネルギー値は扱っていない）。ちなみに、行動決定において Q 学習などを使用する場合には、上式を変換することで、その導入を実現する（適応振動子法による Q 学習法の表現については付録参照）。

$$g \left(\sum_n k_n \frac{d^n}{dt^n} y^n \right) + f(I) = e \quad (1)$$

ここで、

- y : この方程式によって求めるべき解。具体的にはモータの出力角度など。
- k_n : y の n 回微分値に対する係数。
- $g(\quad)$: $\sum_n k_n \frac{d^n}{dt^n} y^n$ の値全体に何らかの影響を与える関数。
- I : 外界からエージェントへの様々な反作用要因の集合である入力ベクトル。
- $f(\quad)$: 入力ベクトル全体に何らかの影響を与える関数。
- e : エージェントの内部エネルギー。

3. 実験設定

次に、提案した設計概念の具現化を行い、行動獲得実験に向けた各種の設定を行う。

3.1 自律分散遊泳ロボット

ハードウェア構造の決定段階において、エージェント間の結合状態はロボットの動作能力に大きく制限を与える。本論文ではエージェント同士を一次的に連結するものとした。以下に自律分散遊泳ロボットとして実際に構築されたものの仕様を説明する。

Fig. 4 の A, B に示すように、本機は円形のフロート（直径 15 [cm]）を一つのエージェントとしており、本論文では三つによる構成を扱う。全長は 80 [cm]、全重量は 5 [kg] である。外界のセンシングには太陽電池を光センサとしてフロート上に四方を向けて配置し、プロセッサとしてはフロート内にワンチップマイコンボード（V55, 16 ビット CPU, NEC 製）を搭載。行動のためのアクチュエータとしては角度制御サーボモータ（S9204,

トルク 9.5 [kgcm]、双葉電子工業製）を、隣のエージェントと連結されるリンクの一端に設け、その部分をリンクの回転軸とした。エージェント間の情報交換のための通信手段としては両隣用に 2 ch のシリアル通信を用いる。本機の移動力はモータの回転によりフロート下部に固定されたフィンが水をかくことで発生される。各エージェントの動きが適切な進行波を生成すれば移動が可能となるが、不適切であればほとんど移動できない構造である。また、構造上必要なため、本機は末端部にダミー・エージェントを設けている。なお、このダミー・エージェントにはロボット駆動用のバッテリーを搭載している。

3.2 その他の実験設定

●動作環境

ロボットの動作環境には、実環境として水中を設定する。これは、具体的には後の Fig. 7 に示すプールである。ここでは、環境本来に流れはなく、他の動作主体も存在しない。しかし、ロボット自らの動作により流れが生じることや、慣性力が無視できないことなどから、過去と同様の動作でも、そのタイミングにより結果的に振舞いの上で大きな差異を生むといった、行動実現の不安定性や不確実性が顕著に現れる。

●通信範囲

各エージェントの通信範囲は両隣までとする。すなわち、エージェント 1 と 3 ではお互いの状況を知ることができない。

●動作決定における意思の調停

意思の調停は特に行わない。学習過程を通じて協調的振舞いが得られることを期待する。

●行動の同期

問題を簡単にするため、行動実行時には全エージェントは同期をとるものとする。

●要素行動の決定～振動運動～

要素行動は、水環境での移動効率のために、正弦関数を基本とした振動運動とする。各エージェントのモータの実行動作となるこの行動は、適応振動子法から次のように導く。

要素行動の決定のためには外界からの入力とは考慮しなくてもよい。また、エネルギーの流動も考慮しないため、式 (1) において、 $f(I) = 0$, $e = 0$ となり、さらに $g(\quad) = 1$ とすると、式 (1) は以下のように表される。

$$\sum_n k_n \frac{d^n}{dt^n} y^n = 0 \quad (2)$$

式 (2) において、 y を x に置き換え、代わりに左辺全体の値を y とし、さらに k_n に式 (3) を代入すると (m は自然数、 n には無限値まで導入)、式 (4) の正弦関数が得られる (x は

Fig. 4 A, Distributed autonomous swimming robot (3 agents); B, Structure of an agent

Fig. 5 Definition of sensing states

Fig. 6 Definition of element action patterns

モータの位相角, y は振幅となる). この振動運動は具体的には次に示す Q 学習の中での要素行動として, 12 種類の変化を持たせて用いられる.

$$k_n = \begin{cases} 0 & , n = 2m \\ \frac{-(-1)^{(n+1/2)}}{n!} & , n = 2m - 1 \end{cases} \quad (3)$$

$$y = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} \dots = \sin x \quad (4)$$

●学習手法の導入～ Q 学習～

学習法としてはいくつか考えられるが, 本論文では実機への適用性, パラメータ獲得後の解析, および再利用性の良さなどから Q 学習 [8] を用いる. 対象とする問題がマルコフ決定過程に基づくものであれば, 連続的に状態を遷移し, 行動学習が可能となることが保証されている [9].

Q テーブルを作成する. 取りうるセンサ状態 S は, 4 方向のうち最大のセンサ値のものを入力とする 4 状態とする (Fig. 5). 図のように右側のセンサ値が最大の場合, $[S:3]$ と表すものとする. また取りうる要素行動パターン A は, 30 度の振幅を持つ正弦関数を基本とし, 3 種類の振動中心角度と 4 種類の位相の組み合わせによる 12 パターンとする (Fig. 6). 各エージェントの選択が適切であれば, システム全体で前後進・左右旋回が可能である. Fig. 6 のように振動中心角が +30 度で位相ずれのない要素行動パターンが選択された場合, その表記は $[A:3-2]$ とし, このとき, モータは図のような動作を実行する. また, 1 ステップの行動における振動運動の実行時間は経験的に 2 周期とした.

Q 値の設定に当たっては, 近傍のエージェントの情報を導入する. しかしながら状態空間サイズの膨大化を防ぐため, ここでは Q 値の環境状態として, 両隣のエージェントのセンサ状態のみを用いる. このため Q 値は次のように表される (状態数

Fig. 7 Experimental equipment

は 768).

$$Q\left(\left(\begin{matrix} S \\ S_{Agent:left}, S_{Agent:self}, S_{Agent:right} \end{matrix}\right), A\right)$$

Q 値の更新には式 (5) を用い, 行動選択においては式 (6) によるボルツマン探索を行う. ここで S_t, A_t はそれぞれ時刻 t におけるエージェントの状態および行動, $Q(S_t, A_t)$ はそのときの Q 値である. また b は時刻 $t+1$ において選択確率が最大となる行動である. α, γ はそれぞれ学習率と減衰率であり, r_t は報酬値である, T は行動のランダム性を決定する熱パラメータである. 実験では経験的に $\alpha = 0.1, \gamma = 0.95, r_t = 100, T = 4.0$ を用いる.

$$\Delta Q(S_t, A_t) = \alpha(r_t + \gamma \max_b Q(S_{t+1}, b) - Q(S_t, A_t)) \quad (5)$$

$$P(A|S) = \frac{\exp(Q(S, A)/T)}{\sum \exp(Q(S, A)/T)} \quad (6)$$

これらの実験設定は適切ではないかもしれない. 例えばロボットの形態や姿勢は, 2 周期分の行動実行中に刻々と変化するため, そのうちの瞬間を入力とするセンサ状態は, ロボットの状態として扱って良いか判断の難しいところである. しかしながら, こうした不安定な状況における行動学習の可能性を検証するのが本研究の主旨であり, センサ状態の設定の仕方をはじめ, その他すべての実験設定は, 選択されうる手段の一つとして試みるものとする.

3.3 実験設備

Fig. 7 に実験設備を示す. ロボットの動作環境は 2.6 [m] 四方のプールで, 一つの隅には目標物として光源を設置する. プール上方には CCD カメラを固定し, その映像を VTR とトラッキングビジョンシステムに送り, ロボットの行動解析に用いる.

4. 実験

4.1 実験 1: 障害物なし目標物接近

障害物回避問題に取りかかる前にまず, 障害物がない状況での目標物接近行動の獲得実験を行う. これは実環境での実験コストを考慮したもので, 初めから障害物を設定した環境で実験を行うとタスク達成確率が低下し, 学習効率が甚だしく悪化するためである (詳細は 5.3 節). この実験で, 任意の状況から目

標物に到達可能な基本的な動作パターンを獲得し、その後の実験で再利用する。

4.1.1 実験方法

光源と対角に位置したプールの隅をロボットの動作開始位置とし、光源に到達したときのみすべてのエージェントに報酬を与え、この試行を繰り返す。

4.1.2 実験結果

405 試行の実験のうち 150 試行目までの結果を Fig. 8 に示す。これは目標物到達に要した各試行ごとのステップ数の遷移である。図より、初めのうち光源到達のステップ数が不安定であったのが、100 試行を超えたあたりから安定し、平均 20 ステップ前後で目標物に到達している様子がうかがえる。

次に示す Fig. 9 の A は 405 試行の実験の後に得られた動作ルールを再生した場合のロボットの行動軌跡（重心座標）の一例である。図中の角の白丸が光源の位置であり、ロボットは図において、左側から光源に接近している様子が分かる。

Fig. 8 A transition of the number of steps needed to reach the target

また、こうした再生行動に定性的に見られる代表的な動作パターンの流れを Fig. 9 の B に、そしてそれらの動作パターンの選択確率の遷移状況を同じく Fig. 9 の C におおの示す。本論文において動作パターンとは、あるセンサ状態において各エージェントの実行する要素行動の結果生じるロボット全体としての挙動のことを指す。図では最もよく現れる七つの条件における動作パターンを挙げている。この結果から理解できることは、ロボットは初めに目標物到達が可能となった条件 1（エージェント 1, 2, 3 に対するセンサ状態が $[S:3, S:3, S:3]$ ）や条件 2（同様にセンサ状態が $[S:3, S:3, S:2]$ ）における動作パターン（エージェント 1, 2, 3 に対しておおの $[A:3-4, A:1-4, A:1-4]$, $[A:3-4, A:1-4, A:1-3]$ ）をサブゴールとして徐々に異なる姿勢からの目標物到達も可能にしていき、最終的に多くの場合、Fig. 9 の B の矢印の向きに行動と、それに伴う状態遷移を繰り返して目標物に到達する、というように振舞いが定性化された。すなわちこの実験により、基本的な大体の動作パターンが獲得されたことが分かる。

4.2 実験 2：障害物回避

続いて、障害物回避行動の獲得実験を行う。本研究での障害物回避における方針は、環境マップの生成や、正確な障害物の認識といった特別な手段を用いずに、先に得られた基本的な動作ルールを再利用しながらできるだけシンプルな方法で解決することを目標とする。

4.2.1 障害物の設置

まず、障害物を設置し、先に得られた動作ルールによるロボットの挙動を観察し、とるべき方法論を検討する。ここでの障害物はアクリル製のフェンスであり、既存の光センサを使用する上でパーセプチュアル・エイリアシング問題 [10] を回避するた

Fig. 9 A, Replay behavior based on acquired action rules; B, Flows of stable motion patterns in replay behavior; C, Transitions of selection probabilities of the seven motion patterns in Fig. 9-B

Fig. 10 A, Replay behavior based on pre-learned Q-table in Obstacle 1; B, Same behavior in Obstacle 2

め、透明のものを用いる。ちなみに通路幅は 32 [cm] で、ロボットの最小幅の 2 倍程度である。

Fig. 10 の A は一つの障害物設置例（これを障害物 1 とする）である。この場合、先の行動獲得実験により、ロボットは目標物に接近しつつも図の左側から寄る動作パターンを獲得していたため、フェンスに横付けの姿勢になった状態でそれ以上接近が不可能になる。また、Fig. 10 の B はもう一つの例（これを障害物 2 とする）であり、この場合、目標物に左から寄る姿勢が通路への侵入角度として不適切であったため、何度挑戦しても通過できない状況に陥る。いずれの場合も 1,000 ステップ（約 2 時間半）の間動作させたがロボットは目標物に到達できなかった。

4.2.2 提案手法～スイッチング Q～

ここで学習法の改良を図る。ロボットの航行学習のように複雑な問題を扱う場合、問題をいくつかの単純なタスクに分解し、学習則を階層的に構築することで、上位の階層ほど問題を抽象化したレベルで解決する手法が注目されており、従来までに様々なものが提案されている [11]～[14]。このような拡張型の学習手法は集中制御によるシステムでは効果的にはたらくことが実証されているが、分散制御システムにおいては目的関数の設計の仕方にも依存すると考えられるが、確実な保証はない。そのため、ここでは分散制御システムに適用することも考慮した新しい拡張型の学習法を提案する必要がある。また、前提としている実機学習において、その効果を検証するための初歩的な実験として、その仕組みはできるだけ単純なものとする。

上記の検証実験により、先に得られたルールを有効活用しながら、より複雑な状況での行動を効率よく学習するための一般的解決策の一つは、「先に得られたルールで進めるところまで進み、デッドロックに陥った場合、その時点から新たなルールを学習する」方法と思われる。そのため、既得の Q テーブルを用いてもデッドロックを回避できない場合に、新たに一つ追加される Q テーブルに切り換えてデッドロックの回避行動を学習するしくみをスイッチング Q として提案する。

【スイッチング Q】

まずスイッチング Q では、デッドロックの判定基準を以下のように定める。

ロボットの持つセンサによって検知される、ロボットの移動距離あるいはそれに準じた値を示すセンサ値 Sv を、最近の j

ステップの間記録し ($Sv_{t-j}, \dots, Sv_t$)、その時系列的な並びを線形関数 $v(\cdot)$ で近似し、その傾き l を求める。この傾きの絶対値 $|l|$ がある値 a 以下であれば、ほとんど移動量に変化はないとして、これをデッドロックと判定する。

$$l = v(Sv_{t-j}, \dots, Sv_t)$$

$$|l| \leq a : \text{Deadlock}$$

次に、スイッチング Q による行動学習の流れを説明する。環境にはデッドロック状態 D が m 個あり、ロボットはすでに n 個の Q テーブルを獲得していることを想定する。

段階 1) D_m までの移動

ロボットは、あるデッドロック状態（スタート状態も含む） D_{x-1} から、ある一つの Q テーブルを用いることによって、次の D_x まで移動する。この間の移動に十分なステップ数を i_x とすると、 D_m に至るまでのロボットの状態遷移は例えば次のように表される。

$$\begin{aligned} \text{Start} &\rightarrow Q_1(i_1) \rightarrow D_1 \rightarrow Q_2(i_2) \rightarrow D_2 \rightarrow Q_1(i_3) \rightarrow \dots \\ &\dots \rightarrow D_{m-1} \rightarrow Q_x(i_m) \rightarrow D_m \end{aligned} \quad (7)$$

ここで Q_x は D_m に至る際に用いられた $Q_1 \sim Q_n$ の中の Q テーブルである。そして D_m を脱出するために次の段階に入る。

段階 2) 既得 Q テーブルの試行・評価

D_m の状態から、それまでに得られた Q テーブルを、ランダム、あるいは Q_1 から順にあるステップ数の間試行し、「 D_m を効率よく脱出可能」、「脱出は可能だが効率は悪い」、「脱出不可能」などのような評価を行う関数を $u(\cdot)$ とし、各 Q テーブルを評価する。

段階 3) D_m 脱出用 Q テーブルの選択

次に D_m を脱出するための Q テーブルの選択を行う。 D_m 脱出用として、ある Q テーブル Q_k が選ばれる確率は段階 2 で行った評価に比例し、この選択確率に従って Q テーブルが選ばれる。

$$Pr(Q_k | D_m) \propto u(Q_k)$$

段階 4) 新規の Q テーブル Q_{n+1} による学習

段階 3 で、ある評価以上の Q テーブルが選択されなかった場合、新たに学習用の Q テーブルとして Q_{n+1} を追加し、 D_m を初期状態とする行動学習を行う。

行動学習時には、時刻 t において用いられる Q テーブル Q_t は、デッドロック判定によって移動がある程度可能とされる場合は、 D_m に至る際に用いられた Q_x を使い、デッドロックと判定される場合は Q_{n+1} により学習を行うものとする。

$$Q_t = \begin{cases} Q_x(S_t, A_t), & |l| > a \\ Q_{n+1}(S_t, A_t), & |l| \leq a \end{cases}$$

学習の際に常に学習用の Q_{n+1} を使うことをせず、一緒に既得の Q_x も用いるのは、実機による学習効率のため、できるだけロボットの行動範囲を D_m の生ずる付近にとどめ、無制限期間のランダム探索による冗長な状態遷移過程の生成を避けるためである。

Q_{n+1} が用いられた場合、そのときの Q 値はいちばん最近に

使われたルール $Rule_{recent}$ として記録される。

$$Rule_{recent} = Q_{n+1}(S_t, A_t)$$

学習において、Q 値の更新は Q_{n+1} に対してのみ行われる。つまり、 Q_{n+1} が用いられる度に $Rule_{recent}$ が更新され、 D_m を回避し、ゴールに到達した際に、このルールに対し報酬が返される。そしてその報酬は式 (4) によって徐々に Q_{n+1} 内に伝搬される。

段階 5) 繰り返し学習

それまでの学習経緯により、状態遷移式 (7) のような、使われた Q テーブルや i_m のステップ数の履歴データは記録済みであるとする。段階 4 において Q_{n+1} が設けられた場合は、これによりロボットはスタート状態から D_m までスムーズに移動し、 D_m からの Q_{n+1} による行動学習を、ある程度の収束が確認されるまで繰り返す。【以上】

このように提案するスイッチング Q について、その有効性を検証する必要があるが、本論文ではまず、その最も基本的なケースとなる $m=1, n=1$ の場合に関して実験を行う。ちなみにここでは経験的に $i_1=9, j=5$ としている。

デッドロック状態や既得 Q テーブルが複数存在するスイッチング Q を実現する場合、段階 2 において、実際に試行を行うのであれば、異なる Q テーブルの試行の開始ごとに、ロボットを確実に D_m にセットし直す何らかの機構が必要となる。そうではなく、実際の試行は行わずに D_m においてはロボット内部で既得 Q テーブルの評価試行を実施することも可能である。この場合、AR モデルなどによるロボットの状態予測を、既得 Q テーブルによる状態遷移シミュレーションと併せて行うことで評価したり、あるいは Dyna [15] などの方法も利用可能であろう。こうした方法で評価された Q テーブルを段階 3 の選択過程に進めればよいと思われる。

本手法は実機学習を前提とするため、方法論はさほどシビアな構成とはなっていない。また、過去の経験を積極利用することを基本方針としているため、新規のルールは既得のルールの不得意な部分を補足的に学習することとなり、最適解の探索がなされにくい点は短所として考えられる。これらの点に加え、本論文では $m=2$ 以上の環境においては、性能評価を行っていないため、今後、実装方法をさらに詳細に決定し、実際に検証する必要がある。

4.2.3 分散型スイッチング Q に期待される特徴

分散システムに適用されるスイッチング Q では、段階 4 におけるデッドロック判定において、エージェントごとに異なる判定もありうる。これによりシステム内では同時期に二つの Q テーブルが行動決定に用いられ、混合された動作パターンの実行が起こり得るが、分散システムの自由度を保つため、敢えてシステム全体として判定を一致させることはしない。

ところが逆にこの設定が、行動学習の上で効率化をもたらすことが期待される。これは、新規のルールの学習と既得のルールの再生が同時期に行われることに起因する。つまり、あるエージェントが既得の効果的な動作パターンを実行する一方で、他のエージェントは新たな動作ルールを学習することで、システム全体として新たな動作パターンを効率よく獲得する効果が期

Fig. 11 Transitions of the number of steps needed to reach the target in Obstacle 1 and 2

Fig. 12 Replay behavior based on Switching Q in Obstacle 1 and 2, and transitions of chosen action rules in the replay tests

待される。

4.2.4 実験方法

先の実験と同様に、光源の対角位置ををロボットの動作開始位置とし、スイッチング Q により、光源到達学習の繰り返し試行を行う。

4.2.5 実験結果

Fig. 11 に、障害物 1 と 2 の場合の目標物到達に要した各試行ごとのステップ数の遷移を表す。始めのうち、目標物に到達するために障害物 1 では 400 ステップ以上、障害物 2 においては 700 ステップ以上要する場合もあったが、次第に少ないステップ数でタスク達成が可能となり、130 試行以降ではどちらも 30 ステップ程度でタスク達成を可能としている。

次に **Fig. 12** として、スイッチング Q による学習後の再生行動の結果を障害物 1 と 2 について 3 例ずつ示す。ここから分かるように、目標物に到達する行動シーケンスは部分的に一致するものもあるが、完全に一致することはあまりなく、多様なパターンが見られる。**Fig. 12** の A では、ロボットは比較的真無駄のない行動シーケンスにより通路を通過しているが、**Fig. 12** の B ではいくらか遠回りして目標物に到達している。また障害物 2 の場合においても、**Fig. 12** の D では一度フェンスに横付けの状態になった後大きく回り込んで通路への侵入を果たしているが、**Fig. 12** の E では、スムーズに目標物に到達している。この結果からいずれにしても 4.2.1 項にて見られなかった障害物回避が達成されていることが分かる。

それから、これらの振舞いにおける各エージェントのとした動作ルールの遷移を、おのおのの図の下に示す。図の中で、濃い色で示されているのが過去に得られたルールで、薄い色で示されているのが新たに得られたルールである。スイッチング Q の設定により、どのエージェントも初めの 9 ステップは無条件に既得のルールを選択し、それによりフェンス付近まで接近しているが、その後は各エージェントのデッドロック検知に基づいて個別に新たなルールと既得のルールをスイッチしている様子が確認できる。

5. 議論と解析

得られた結果をもとに、分散制御システムにおける行動学習のメカニズムの解析に焦点を絞って議論を行う。

5.1 分散制御による行動学習について

まず、単純な目標物接近行動の獲得の仕組みを考える。ここで議論すべきは、エージェント単位では局所的情報のみに依存した行動学習であるにもかかわらず、なぜ結果的にシステム全体としての動作パターンが得られるのか、という点である。各エージェントはシステム全体の状態を知ることができないが、システム内には何らかの方法で未知の情報を外挿するような仕組みがはたらいっていると考えられる。

適応振動子法の設定では、各エージェントは近傍のエージェントの状態までは局所情報として知ることができる。そのため、すべてのエージェントが近傍の状況を知ること、情報共有のための弱いネットワークがエージェント間に形成されることが期待される。一方で、Q 学習のはたらきにより、各エージェントごとに動作シーケンスが学習されることが期待される。これ

Table 1 Rates of reused rule transition in replay experiments

$Q(S_t, A_t)$ $Q(S_{t+1}, A_{t+1})$	single Q-learning		Switching-Q	
	initial phase	non obstacle	Obstacle 1	Obstacle 2
Rate of Inter-agent (%)	7.3	90.5	64.6	63.6
Rate of Inner-agent (%)	17.1	91.0	68.1	65.9

によりシステムは、エージェント間あるいは時系列間において、状態の暗黙的な結合関係を形成する可能性を持つことが考えられる。また加えて、協調的な動作のための同一の目的関数設定や、機械的結合という構造により、あるエージェントの振舞いが他のエージェントの行動学習に影響を与えることも考えられる。

以上に推測されるはたらきは、実際に **Fig. 9** の C に見ることができる。図からは、早い段階に選択確率を上げたエージェントの要素行動パターン（例えば条件 3 におけるエージェント 3 の要素行動パターン [A:1-2]）が基盤となって、それに続く動作パターン（条件 4 や 5 の動作パターン）の選択確率が上がる様子が観察される。

こうした行動学習過程を通じ、獲得されたルールの定量的な再現性を調査するために、行動の再生実験においてルールの遷移過程が同じになる割合を計算した。**Table 1** は 10 回の再生試行におけるその統計データである。この中で「Rate of Inter-agent」とは、隣り合う二つのエージェント間で、時刻とにおいて用いられたルールの組にまったく同じ組み合わせが現れたパーセンテージを表すもので、一方「Rate of Inner-agent」とは、同一のエージェントにおいて同様のパーセンテージを表すものである。この表から、単純な目標物到達学習である単一の Q 学習の場合について見ると、学習の初期段階では双方の場合とも 20% を切る低い割合であるのが、学習後には 90% 以上にまで高くなっていることが分かる。すなわち、分散学習の結果、特定の動作パターンの選択確率が高くなるように、隣接したエージェント間や同一エージェントの時系列上の動作の中に、強い結合関係が暗黙的に形成されたことが分かった。そのため最終的に、システム全体としてシーケンシャルな動作が再現性をもって実現可能となることが理解できる。

5.2 障害物回避行動獲得のメカニズムについて

次に、障害物回避行動の獲得のしくみについて解析を行う。議論点は、スイッチング Q のはたらきであり、スイッチング Q の導入により、各エージェントは何を学習したのか、という部分である。

これを確認するために、二つの障害物設定において、スイッチング Q の中で新規に導入された Q テーブルのみを用いた再生行動を観察する。**Fig. 13** の A は障害物 1 の場合であるが、ここでは、多少時間がかかるがフェンスの正面に向かうことなく、直接下側から回り込むように通路を通過する振舞いが確認された。また一方、障害物 2 における **Fig. 13** の B では、フェンスの右下側で停滞してしまう振舞いが見られた。**Fig. 10** の A、B と見比べても分かるように、これらは初めの Q テーブルに基づく振舞いとはまったく異なるものである。つまり新規の Q テーブルは、初めの Q テーブルと組み合わせるか、あるいは場合に

Fig. 13 A, Replay behavior based on new Q-table in Obstacle 1; B, Replay behavior based on new Q-table in Obstacle 2

Fig. 14 Motion patterns which became conclusive evidence for gate entry; A, 20th motion pattern in Fig. 12-C; B, 34th motion pattern in Fig. 12-D

よってはそれだけを用いることによって目的行動を実現するように学習されたと考えられる。

この推察を裏づけるために、各障害物設定において通路に侵入する決め手となった行動をクローズアップしてみる。Fig. 14のAはFig. 12のCで通路侵入に至った20ステップ目の行動で、[S:2, S:1, S:1]において[A:1-3, A:2-3, A:2-2]という、左に旋回しながらの前進である。これは、Fig. 9のCの条件3と同じセンサ状態であるが、動作パターンが通路侵入に不適切であったためか、すべて初めのQテーブルに基づく要素行動とは異なる、新たな動作パターンが獲得されている。一方、Fig. 14のBはFig. 12のDで通路に侵入した34ステップ目の行動で、[S:2, S:3, S:2]において[A:2-1, A:2-3, A:3-2]という、通路への侵入角度が適切な行動である。この場合、初めのQテーブルと新規のQテーブルの組み合わせにより障害物回避を実現している。すなわちこの結果から、各エージェントが状況に応じて適切にQテーブルを使い分けるように行動獲得を行ったことが確認された。

さらに、スイッチングQの学習結果の定量評価を、先程と同様に、再現行動での使用ルールの組み合わせが同一となる割合を算出することで行う。Table 1において、障害物回避学習であるスイッチングQの場合について見ると、すべて65%前後の値となっている。本論文でのスイッチングQの場合、Qテーブルは二つで、単一のQ学習時に比べ、単純にルール数が2倍となっていることに加え、分散システムのために二つのQテーブルの切り替えは非同期に行っている。それにもかかわらず、システム全体としての実行行動の再現性が65%を超えるこ

とは、ルール間に強い結合関係が形成されていることを表しており、それゆえ障害物回避行動が実現されていることが分かる。

5.3 分散スイッチングQ学習の有効性

分散制御システムにおけるスイッチングQによる学習時間の効率を調べる。障害物の1と2において、単一のQテーブルの初期状態（行動の選択確率がすべて0）から行動獲得実験をおのの行ったが、どちらの場合も1試行目から1,000ステップ以上経過しても目標物に到達できず、数回繰り返したもののまったく不可能なため、現実的ではないと判断し、実験を中断した。

この結果に見られる実時間学習の可能性に加え、先に記した成果に見られる分散エージェント間での協調学習の推進性や既得ルールの有効利用性などから、提案したスイッチングQは分散制御システムのための拡張型学習法として十分有効であることが確認された。

6. おわりに

実環境で適応的に行動パターンを生成可能なロボットの実現を目的として、機械的に結合された自律分散システムによるマルチエージェントロボットを設計した。そこでは、外部環境や他の構成要素との積極的な相互作用によって動作パターンを獲得することを特徴とした適応振動子法を提案し、その他に、同一の目的関数の設定や、エージェントの同一構成などを条件とした。こうした設計概念に基づくロボットによる動作パターン獲得の検証のため、構築された自律分散遊泳ロボットを用いて、水上での目標物接近と障害物回避の行動獲得実験を行った。学習法にはQ学習を導入し、障害物回避においては既得の動作ルールの有効活用を考慮したスイッチングQ学習を提案した。実験の結果、ロボットは簡単な障害物回避行動を獲得し、獲得されたルール間には時系列的な結合関係が確認され、提案した行動学習法が有効であることが検証された。今後は、より複雑な環境において行動学習の実験検証を行い、提案方法のさらなる理論展開を目指す。

参考文献

- [1] 伊藤, 市川, 須田: 自律分散宣言. pp.2-14, オーム社, 1995.
- [2] A. Cai, T. Fukuda, F. Arai, H. Ishihara, K. Yamada and S. Matsumura: "A Study on Dynamically Reconfigurable Robotic Systems (28th Report, Sharing of Sensing Information in Cellular Robotic System)," Trans. of the Japanese Soc. of Mech. Eng. (JSME), vol.63, no.611, pp.206-213, 1997.
- [3] 村田, 黒河, 小鍛冶: "自己修復する機械", 計測自動制御学会論文集, vol.31, no.2, pp.254-262, 1995.
- [4] M. Yim: "A Reconfigurable Modular Robot with Many Modes of Locomotion," Proc. of the JSME Int. Conf. on Advanced Mechatronics, pp.283-288, 1993.
- [5] D. Iijima, W. Yu, H. Yokoi and Y. Kakazu: "Autonomous Acquisition of Adaptive Behavior for a Distributed Floating Robot Based on the AHC Method," Intelligent Engineering Systems through Artificial Neural Networks, vol.8, pp.537-542, 1998.
- [6] 飯島, 兪, 横井, 嘉数: "分散制御型遊泳ロボットの自律的な目標物接近行動の獲得", 日本機械学会論文集 (C編), vol.65, no.673, pp.208-215, 1999.
- [7] D. Iijima, W. Yu, H. Yokoi and Y. Kakazu: "Distributed Robotic Learning: Adaptive Behavior Acquisition for Dis-

tributed Autonomous Swimming Robot in Real World," Machine Learning, Proceedings of the Sixteenth International Conference (ML '99), pp.191-199, 1999.

- [8] C.J.C.H. Watkins and P. Dayan: "Technical Note: Q-Learning," R.S. Sutton (ed.), Reinforcement Learning, pp.55-68, 1993.
- [9] L. Christman and M. Littman: "Hidden State and Short-Term Memory," Presentation at Reinforcement Learning Workshop, Machine Learning Conference, 1995.
- [10] R.A. McCallum: "Instance-Based Utile Distinctions for Reinforcement Learning with Hidden State," Machine Learning, Proceedings of the Twelfth International Conference (ML'95), pp.387-395, 1995.
- [11] L.J. Lin: "Scaling Up Reinforcement Learning for Robot Control," Machine Learning, Proceedings of the Tenth International Conference (ML'93), pp.182-196, 1993.
- [12] S.P. Singh: "Scaling Reinforcement Learning Algorithms by Learning Variable Temporal Resolution Models," Machine Learning, Proceedings of the Ninth International Conference (ML'92), pp.406-415, 1992.
- [13] M. Wiering and J. Schmidhuber: "HQ-Learning," Adaptive Behavior, 6, 2, pp.219-246, 1997.
- [14] J. Tani and S. Nolfi: "Learning to Perceive the World as Articulated: An Approach for Hierarchical Learning in Sensory-Motor Systems," From Animals To Animats 5 (SAB 98), pp.270-279, 1998.
- [15] R.S. Sutton: "Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming," Proceedings of the Seventh International Conference on Machine Learning (ML'90), pp.216-224, 1990.

付 録

【適応振動子法による Q 学習法の表現について】

2 章で紹介した適応振動子法の式 (1) を以下に記す。

$$g\left(\sum_n k_n \frac{d^n}{dt^n} y^n\right) + f(I) = e \quad (1)$$

適応振動子法では入力する外界因子として、入力集合 I を定めており、この I は、時刻 t において観測される自分の状態 S_t^{self} と近傍エージェントの状態 $S_t^{neighbor}$ からなる状態次元 S_t 、および外界から得られるエネルギー次元 E とから構成されるものと定義している [5]。すなわち、

$$I = \{S_t = \{S_t^{self}, S_t^{neighbor}\}, E\}$$

である。

しかし、適応振動子法から本論文のように Q 学習を表現する場合、以降のような若干の変更を必要とする。つまりこの場合、適応振動子法において解、あるいは出力値としている y を、単なるスカラー量ではなく、内部に状態次元 S_t や行動次元 A_t を持ち、それ自体が時間変化に伴って評価を行うような関数 $V(\cdot)$ として考える。すなわち、

$$y = V(S_t, A_t)$$

である。

本論文では外部環境とのエネルギー授受は考慮していない。また、状態次元は評価関数 y に含まれるため、実際には式 (1) において、 $f(I) = 0$ 、 $e = 0$ となり、さらに $g(\cdot) = 1$ とする

と、式 (1) は以下のように表される。

$$\sum_n k_n \frac{d^n}{dt^n} y^n = 0 \quad (8)$$

式 (8) を展開すると、

$$\begin{aligned} \sum_{n=1}^{\infty} k_n \frac{d^n}{dt^n} y^n &= k_1 \frac{y_1 - y_{t-1}}{dt} \\ &\quad + k_2 \frac{\frac{y_t - y_{t-1}}{dt} - \frac{y_{t-1} - y_{t-2}}{dt}}{dt} + \dots \\ &= l_0 y_t + \sum_{n=1}^{\infty} l_n y_{t-n} \quad (l_0, l_n \text{ は任意の係数}) \end{aligned} \quad (9)$$

ここで式 (9) を整理すると、次の式 (4) が得られる。

$$y_t = \sum_{n=1}^{\infty} m_n y_{t-n} \quad (m_n \text{ は任意の係数}) \quad (10)$$

式 (10) の意味するところは理想的な AR モデルと同値であり、時刻 t における評価関数 y_t の決定のために、理論上、過去の無限ステップまでの加重和をとることで算出している、というものであり、様々な非線形関数を表現しうるような時系列予測の枠組みとなっている。

ところで、強化学習においては評価関数は長期報酬の和を最大化することが目的とされ、 r を報酬、 γ を割引率として、最近の報酬ほど高い割合で重み付けされるように、一般に次のように計算される。

$$V_t = \sum_{n=0}^{\infty} \gamma^n r_{t+n+1}$$

これを展開し、整理すると、

$$V_t = r_{t+1} + \gamma^n \sum_{n=0}^{\infty} \gamma^n r_{t+n+2} = r_{t+1} + \gamma V_{t+1}$$

ここで、 V_t の実現の一つの例として状態 s_t と行動 a_t のセットに対し評価するような関数 $Q(s_t, a_t)$ を考えた場合、 $Q(s_t, a_t)$ の理想値と実際の値との間の誤差分 $\Delta Q(s_t, a_t)$ は、さらに学習率 α を掛け合わせるによって以下のように表される。

$$\Delta Q(s_t, a_t) = \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \quad (11)$$

式 (11) は漸化的に見ることで、理論的には長期的に評価予測をしていることと同等に考えられるが、外見的には 1 ステップ更新の式であり、式 (10) の特殊なケースとして考えることができる。そしてそのはたらきとしては、Q 値の学習によって結果的に行動と状態のセットに対し、時系列関係を学習することを可能としている。更に実装段階において、Q 学習では、評価関数 Q の更新を式 (11) に基づいて行い、行動選択においてはソフトマックス的な手法を用いることでこれを実行する。

以上の説明から，概念的な意味合いにおいて適応振動子法は十分に Q 学習法を表現しうるということがいえる。

飯島大典 (Daisuke Iijima)

1973 年 7 月 31 日生。1998 年北海道大学大学院工学研究科システム情報工学専攻修士課程修了。現在，同大学大学院博士課程に在籍。自律分散システム，自律ロボット，適応行動システムの研究に従事。日本機械学会，ISAB の会員。
(日本ロボット学会学生会員)

横井浩史 (Hiroshi Yokoi)

1963 年 8 月 1 日生。1993 年北海道大学大学院工学研究科精密工学専攻博士課程修了。通産省工技院生命工学工業研究所勤務などを経て，現在，北海道大学工学部システム情報工学専攻助教授。人工知能，人工生命，強化学習，などの研究に従事。博士（工学）。日本機械学会，計測自動制御学会，IAKE などの会員。
(日本ロボット学会正会員)

俞 文偉 (Wenwei Yu)

1967 年 6 月 3 日生。1997 年北海道大学大学院工学研究科システム情報工学専攻博士課程修了。現在，同大学工学部システム情報工学専攻助手。人工知能，ロボット制御，ロボット学習に関する研究に従事。博士（工学）。

嘉数侑昇 (Yukinori Kakazu)

1941 年 10 月 3 日生。1975 年北海道大学大学院工学研究科精密工学専攻博士課程修了。旭川工業高等専門学校助教授，教授などを経て，現在，北海道大学工学部システム情報工学専攻教授。機械学習，進化的計算理論，人工生命などの研究に従事。工学博士。1990 年精密工学会技術賞，1994 年 JSME 功労賞受賞。日本機械学会，計測自動制御学会，などの会員。
(日本ロボット学会正会員)