様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　２６　日現在

機関番号：　５８００１
研究種目：　若手研究(B)
研究期間：　2013～2016
課題番号：　２５８７１０４９

研究課題名（和文）強化学習個体群における行動時系列を基にしたコミュニケーション創発メカニズムの解明

研究課題名（英文）Elucidation of communication emergence mechanism based on action time series in reinforcement learning agents.

研究代表者

佐藤　尚（Sato, Takashi）

沖縄工業高等専門学校・メディア情報工学科・准教授

研究者番号：７０４２６５７６

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：ジェスチャー理論を基に原始的コミュニケーションの創発に必要な個体の能力とその他の要素について議論した。個体の能力面を検証するため、衝突回避ゲーム、およびそのプレーヤとして行動履歴を学習する強化学習個体を採用した。実験の結果から、過去の行動履歴情報をその順番も含めて学習・予測する能力がコミュニケーションの創発に重要な役割を果たすことが示唆された。また、コミュニケーションの成立に寄与しうる要素を検討するため、拡張版SOM学習個体によるコミュニケーションゲームを採用した。この実験より、個体以外の状況から得られる暗示的フィードバックがコミュニケーション成功率を向上できる可能性があることが示唆された。

研究成果の概要（英文）：Based on the gesture theory, we discussed an individual's ability and other factors necessary for emergence of proto-communication in a primitive society in which the communication was not established among the individuals. To verify the individual's ability aspect, we adopted a collision avoidance game and a reinforcement learning agents who can learn their action history as the game players. Our simulation showed that, by evaluating various models including a hybrid model between the Q-learning and the recurrent neural network, the abilities to learn and predict the past action history and its order can be played an important role in the emergence of communication. Also, to examine an element contributed to the formation of communication, we adopted a communication game with extended SOM learning agents. The second simulations suggested that "implicit feedback" obtained from situations other than individuals, which is proposed by us, can be improved the communication success rate.

研究分野：　複雑系、人工生命、進化言語学、進化的計算論

キーワード：　ジェスチャー理論　原始的コミュニケーションの創発　Q-learning　Neural Q-learning　Recurrent Q-learning　マルチエージェント・システム　拡張版SOM　暗示的フィードバック

１．研究開始当初の背景

言語の起源と進化の問題を主に扱う進化言語学[橋本, 2004]の立場からの研究では、理解したい対象の基となる数理モデルを作り、それをコンピュータシミュレーションやロボット実験により動かすことを通してその対象の理解を試みる構成論的手法[Kaneko&Tsuda, 2000]を用いることによって、言語の起源と進化に関する様々な知見が得られており、近年興隆を見せている。この手法を用いた言語やコミュニケーションの起源に関する研究の多くは、記号接地問題[Harnad, 1990]への記号主義的[Vogt, 2006]ないし分散主義的[Cangelosi, 2006]アプローチによるものであり、ソフトウェアエージェントやロボット間での相互作用や行動の学習による、文字列等で表された記号コミュニケーションの創発について議論したものばかりである。

これらの殆どが、基本的には「音声」を言語の起源として想定したものであると考えられる。しかしながら、初期人類の口腔・咽頭の構造が多様かつ複雑な音声を発するのに向いていなかった点、そもそも音声が自由自在に制御できるものではなく、状況等に対応した反射的なものであったと考えられる点など、言語の起源として考えるにはいくつかの問題があることが指摘されている。一方、音声と比較して手足は意図通りに、かつ正確に制御でき、それらの基礎的な「動作」によって表される「ジェスチャー」はコミュニケーションを成立させる上で「意味」を載せる媒体として有効なものであったと考えられる[Tomasello, 1996][Corballis, 2002]。

このような基礎的行動に基づくコミュニケーションの創発を扱ったものとして、高野らの研究[高野&有田, 2006]がある。この研究では、複数のエージェントが衝突を回避しながら各自に与えられたゴールに到達することを目的としたゲームを用いた進化シミュレーションを行なっている。この進化シミュレーション実験により、コミュニケーションが成功したことに報酬を与えていないにも関わらず、基礎的行動を用いて回避方向を知らせるという協調行動が創発することを示した。しかし、実際のヒトのコミュニケーションの創発を想定する場合、コミュニケーションの成立に必要となる知識や方法は先天的に持つのではなく、学習によって後天的に獲得されるものだと考えられる。

ここで、高野ら[高野&有田, 2006]がエージェントの行動決定システムとして採用した Recurrent Neural Network (RNN) の結合重みを進化ではなく学習で変化させることを考える場合、一般的に RNN で用いられる学習アルゴリズムでは、その RNN に出力してほしい理想的な解(の時系列)が必要である。ところが、基礎的行動、またはその組み合わせを「ジェスチャー」として利用したコミュニケーションを成立させられるようになるかどうかを検証しようとするならば、どのような行動をすれば良いのかということを予め決めておくわけにはいかない。このことから、試行錯誤的に様々な行動を試しながら、各状況に適した行動を選択できる学習法の 1 つである「強化学習」が用いられるべきだと考えられる。しかし、従来の強化学習単体では時系列データの学習が難しい。このような問題を解決するために、強化学習と時系列処理が可能である RNN を組み合わせたモデルが提案されている[Lin&Mitchell, 1993]。しかしながら、上述の通り、時系列学習が可能な強化学習法を採用した、「ジェスチャー」が言語の起源であるとする仮説に基づくコミュニケーション創発に関する研究は殆ど見当たらない。

２．研究の目的

本研究では、原始的コミュニケーションがジェスチャーで行われていたという言語の起源に関する仮説に基づき、基礎的行動を強化学習法によって学習するエージェントがコミュニケーションとは関係ない目的を持った「行動」をどのようにしてコミュニケーション上の「記号」として用いるようになるのか、またそれは、どのような能力を持つ個体間において、どのような要素や条件が必要であるのかを明らかにすることを目的とする。

３．研究の方法

本研究では、高野ら[高野&有田, 2006]が用いていた「衝突回避ゲーム」を採用する。しかし、高野らの研究と大きく異なる点は、彼らがエージェントのモデルとして採用した RNN の一種である「エルマンネット[Elman, 1990]」を進化論的手法ではなく「学習」で調整する点にある。基礎的行動の試行錯誤的学習を通して、その行動をコミュニケーション上の「記号」に利用できるようになるのかを検証する。

具体的には、図１に示されるように、２次元連続空間において、複数のエージェントとそれらのためのゴールをランダムに配置する。各エージェントは扇型の視界を持つ対向２輪の車輪駆動型とし、ゴールまでの距離、自分の視線の角度、視界内の最近傍個体との距離とその個体の視線の角度などの情報を得ながら自分のゴールに到達することを目標として移動する。なお、壁や他個体と衝突した場合、罰を与え、ゴールに近づいた時および到着できた時のみ報酬を与えるものとする。

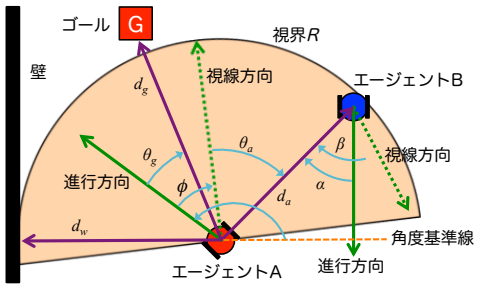

図１：強化学習エージェントが受け取る情報の種類(高野らの論文[高野&有田, 2006]の図を基に加筆修正して作成)。

基礎的行動をコミュニケーションに利用することができるようになる個体の学習能力としてはどのようなものが必要であるかを調べるため、以下の３つの比較研究を行った。

(1) 自分に対しての最近傍個体の進行方向の角度 α と視線方向の角度 β に注目して構

成した 9 種類の Q-learning(QL)エージェント(表 1)の比較

表 1：α と β の入力情報と 9 種類の QL エージェントの関係（×は情報を受理しない）。

モデル	α (t)	α (t-1)	β (t)	β (t-1)
1	×	×	×	×
2	○	×	×	×
3	○	○	×	×
4	×	×	○	×
5	×	×	○	○
6	○	×	○	×
7	○	○	○	×
8	○	×	○	○
9	○	○	○	○

(2) QL エージェントと Neural Q-learning[Kuzmin, 2002] (NQL) エージェント(図2)との比較

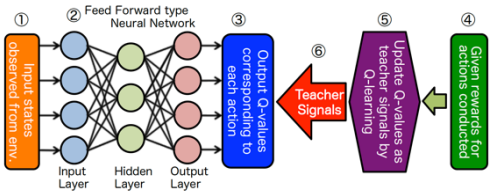

図 2：Neural Q-learning の概念図。

(3) NQL エージェント[Kuzmin, 2002] と Recurrent Q-learning (RQL) エージェント[Lin&Mitchell, 1993] (図3)との比較

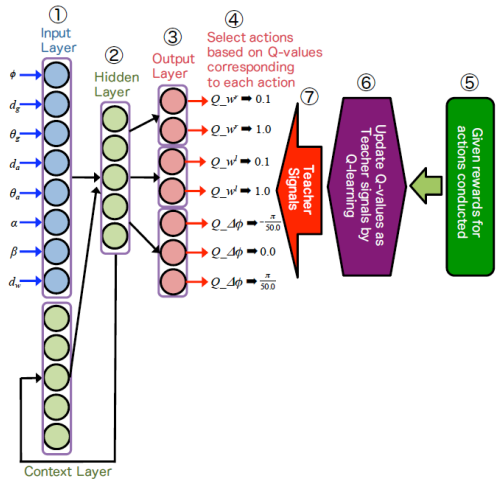

図 3：Recurrent Q-learning の概念図。

また、個体間においてコミュニケーションがまだ成立していなかったような原始社会では、個体同士が直接的にコミュニケーションによってそのコミュニケーションの内容自体の訂正などを行うことができないと考えられる。

このことより、個体以外の「状況」からの間接的なフィードバックがあったのではないかと考えた。これを本研究では「暗示的フィードバック」として提案し、最後に 4 つ目の研究として、この暗示的フィードバックがコミュニケーション形成に寄与しうる要素の 1 つとなるのかを検証した。

(4) 拡張版自己組織化マップ(SOM)学習エージェント(図 4)によるコミュニケーションゲーム[林&有田, 2004]での暗示的フィードバックの効果の検証

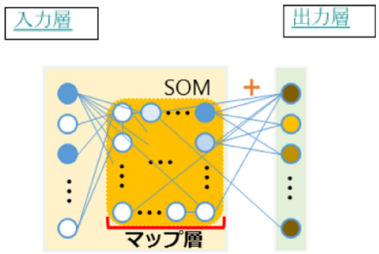

図 4:拡張版 SOM 学習エージェントの概念図。

4 つ目の研究では、原始社会のような、形成される概念が個体ごとに異なり、またコミュニケーションの結果を明示的にフィードバックすることができない個体間でのコミュニケーションについて議論した。このような原始社会的状況において、コミュニケーション・システムの形成に寄与し得る要素について調べるため、明示的なフィードバックを行えない状況でのコミュニケーション創発について検証した林らのモデル[林&有田, 2004] (図 4)に具体的な拡張として以下の 2 つを導入したものを採用する。なお，本研究では、林らの研究とは異なり、「身振り手振り」などの「基本動作」がオブジェクトから形成した概念と結びついて外部に表現されると考え、その「基本動作」を「シグナル」と呼ぶことにする。

・形成される概念が個体毎に異なる環境。
・カテゴリに基づく明示的ではないフィードバック＝暗示的フィードバック。

ここで「カテゴリ」とは、エージェントが環境中で認識する様々なオブジェクトに共通する特徴を基に推測されるもので、モデルの中ではベクトルによって表される。また、本研究では暗示的フィードバックとして次の 2 種類を提案・採用する：

・シグナルに対する暗示的フィードバック。
・カテゴリに対する暗示的フィードバック。

前者は、あるカテゴリに対して複数のシグナルが結びついている状況を想定する。これは、あるカテゴリに対して結びつけているシグナルが、両エージェント間で異なる状況を示唆する。そこで本研究では、このような状況において、ランダムに選択した 1 つ以外のシグナルの重みを下げるフィードバックを行うものとする。

後者は、あるシグナルの解釈ニューロン(出力層のニューロン＝シグナルニューロンと最大の荷重で結合している SOM のマップ層上のニューロン)が複数見つかるような状況を想定する。この状況では、ある個体が別々のカテゴリに分類しているオブジェクトを、その相手が 1 つのオブジェクトとして捉えている状況を示唆する。そこで本研究では、このような状況において、複数の解釈ニューロンからそれぞれ解釈されるカテゴリを探し出し、それらのカテゴリを基に新しいカテゴリを作成するものとする。

最後に、拡張版 SOM 学習エージェントたちがプレイするコミュニケーションゲームの簡単な説明を以下に示す：

1) N_a 体のエージェント群から 2 体をランダムに選択する。

2) 環境中の N_o 個のオブジェクトから N_c 個のオブジェクトとその見え方(特徴値)をランダムに選択し、それをコンテキストとする。
3) 下記手順 4)～7)を 2 回行う。この時、エージェントはシグナル発信者またはシグナル受信者の役割に分かれ、2 回目はその役割を交代する。
4) シグナル発信者とシグナル受信者それぞれがコンテキスト中のオブジェクトに対し、どのカテゴリに属するかを決定する。
5) シグナル発信者がコンテキストからカテゴリをランダムに 1 つ選択し、それをトピックとする。
6) シグナル発信者は SOM にトピックのカテゴリベクトルを入力し、それを解釈できるシグナルを示す。
7) シグナル受信者はシグナル発信者が示したシグナルを解釈し、コンテキスト中のどのオブジェクトについて言及したのかを推測する。
8) マップ層-出力層間の学習を行う。
9) カテゴリの信頼度を更新する。

４．研究成果

(1) α と β の入力情報の違いによる 9 種類の QL エージェントの比較結果

現在時刻の α と β をそれぞれ単独、またはそれら両方を参照して学習する場合、大きなパフォーマンスの向上は見られなかった。しかし、一時刻前の α と β のいずれかを他の情報と一緒に参照することで、衝突回避(図 5)、およびゴール到達能力(図 6)を向上させられることを確かめた。

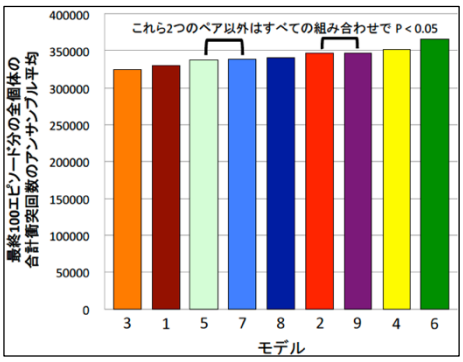

図 5：9 種類の QL エージェントモデル各々の合計衝突回数のアンサンブル平均。

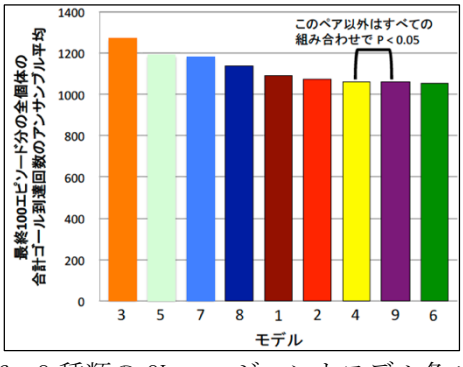

図 6：9 種類の QL エージェントモデル各々のゴール到達回数のアンサンブル平均。

β には予め特別な意味などを与えていたわけではないため、基本的には衝突回避に何の役にも立たない情報であるにもかかわらず、上述の結果が得られたことは非常に興味深い。これらの結果は、高野らの研究[高野&有田, 2006]で示された、多数のエージェントがそれぞれ衝突回避しながらゴールへ到達するという協調行動を「学習」によっても実現できる可能性があることを示すものである。

しかし、本比較研究では、残念ながら、エージェントが現在と過去の β の情報と自身の視線方向を左右に動かすという基礎的行動プリミティブをコミュニケーション上の「記号」として利用して衝突回避を実現したことを示唆する状況証拠的結果を示すのみにとどまっている。

(2) QL エージェントと NQL エージェントの比較結果

エージェントが受理する情報が QL で用いる離散値と NQL で用いる連続値の各々で相互作用にどのような違いが生じるのかを検証した。

衝突回避においては QL エージェント(図 7)、そして、ゴール到達においては NQL エージェント(図 8)が高い性能を持つことが示された。

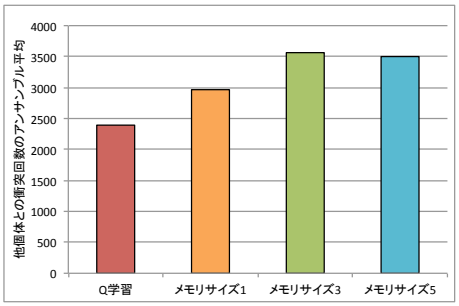

図 7：QL、およびメモリサイズが異なる 3 種類の NQL エージェントモデル各々の衝突回数のアンサンブル平均。

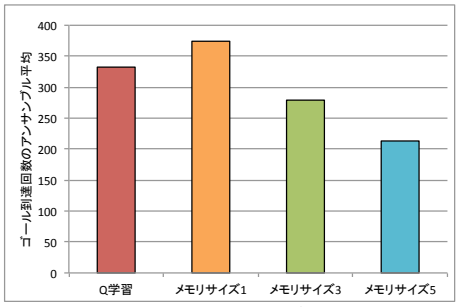

図 8：QL、およびメモリサイズが異なる 3 種類の NQL エージェントモデル各々のゴール到達回数のアンサンブル平均。

これは、衝突回避が他個体についてのおおまかな情報のみでも十分可能であり、ゴール到達にはゴールの角度や位置についての正確な情報を得ることが必要であるからと考えられる。また、本研究で用いた NQL エージェントにおいては、メモリサイズ 1 のエージェントが、衝突回避においてもゴール到達においても最も高い性能を示した。このようになった理由として、NQL エージェントで用いているフィードフォワード型のニューラルネットワークでは

時系列データの時間的構造まで適切に学習することができず、過去情報を増やすことによって学習する内容が複雑になるほど時系列データ間に矛盾が生じやすくなることが原因であると考えられる。

本比較研究においても、具体的なコミュニケーション創発現象を確認することはできなかった。

(3) NQL エージェントと RQL エージェントとの比較結果

前述の(1)と(2)の研究結果より、エージェントが経験した過去から現在までの履歴情報、およびその順序を学習できる能力がコミュニケーションを創発する上で重要な役割を果たし得る可能性があることが示唆された。そこで時系列データを学習可能な RQL とそれが不可能な NQL との比較検証を行った。

RQL エージェントは最大学習回数を増やすことによって、ゴール到達回数が大幅に減少したが(図 9)、他個体との衝突回避のパフォーマンスは逆に向上した(図 10)。

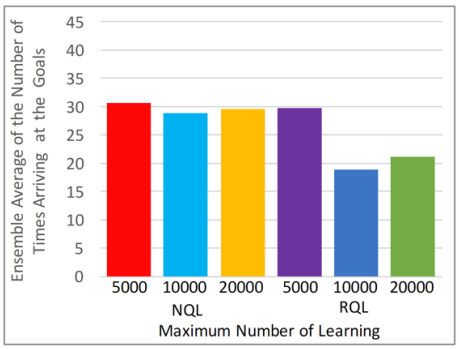

図 9：最大学習回数が異なる 3 種類の NQL、および 3 種類の RQL エージェントモデル各々のゴール到達回数のアンサンブル平均。

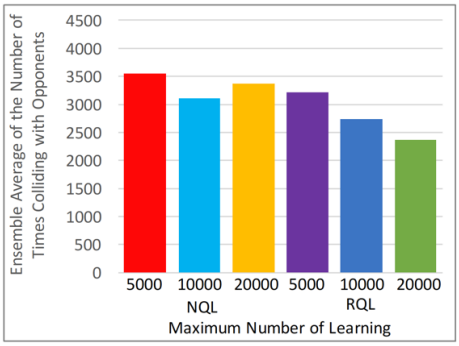

図 10：最大学習回数が異なる 3 種類の NQL、および 3 種類の RQL エージェントモデル各々の衝突回数のアンサンブル平均。

本ゲームでは、高い頻度で衝突が発生するため、学習プロセスを通じて優先的に衝突を回避する能力を得たのではないかと考えられる。しかし、RQL エージェントの価値関数を適切に変化させられるだけの学習とエピソードの最大回数を十分に大きくした場合、ゴール到達、および衝突回避の性能の両方を向上させる可能性はある。

また、稀ではあるが、RQL エージェントが視線を移動させるという基礎的行動を衝突回避のための記号として用いる原始的コミュニケーション・システムが創発することが分かった(図 11)。

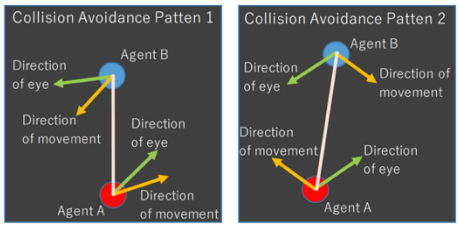

図 11：視線を動かすという基礎的行動を利用して他個体との衝突回避を実現する 2 種類のコミュニケーションの創発。

更に、最大学習回数が多い場合には、RQL エージェントが他個体と壁との衝突を避けるために同じ場所で回転し続ける行動を獲得できることも確認した(図 12)。特に、RQL エージェントの時系列学習能力は、自身の回転運動の回転数をカウントして他個体が通過するまで待つ、ということを実現する上で重要な役割を果たすと考えられる。

高野ら[高野&有田, 2006]の研究よりもコミュニケーションの創発を確認できなかった理由としては、進化と学習の違いが挙げられる。前者の場合、適応的な遺伝子が広まるため、同じコミュニケーション・システムを最初から備える個体が作られ易くなる。一方、学習では各々が独自のシステムを獲得するため、それを共有化することは難しい。しかし、非常に長い時間を掛けて学習させ続けることで、他者との相互作用を通じて、互いに自身の持つコミュニケーション・システムを学び合い、共有化できる可能性があることが示唆された。

以上より、RQL エージェントが持つ時系列学習・予測能力は、ジェスチャーによるコミュニケーションの創発に必要な個体の能力の 1 つである可能性があると考えられる。

(4) 拡張版 SOM 学習エージェントによるコミュニケーションゲームでの暗示的フィードバックの効果の検証結果

暗示的フィードバックには，各エージェントの SOM の構造が大きく異なっている場合、あるいは異なるオブジェクトを同じだと認識してしまう概念を形成している状況においても、コミュニケーションの成功率を向上させる働きがあることが分かった(表 2)。

表 2：SOM の 3 種類の学習回数での各モデルの平均コミュニケーション成功率((1)カテゴリを導入したモデル、(2)上記(1)にシグナルに対する暗示的フィードバックを導入したモデル、(3)上記(1)にカテゴリに対する暗示的フィードバックを導入したモデル、(4)上記(1)に 2 種類の暗示的フィードバックを導入したモデル)。

モデル	$T_{som}=(30\times N_o)$	$T_{som}=(300\times N_o)$	$T_{som}=(3000\times N_o)$
(1)	0.260439	0.380189	0.392413
(2)	0.271754	0.400125	0.434301
(3)	0.266199	0.397116	0.471118
(4)	0.294536	0.427935	0.498801

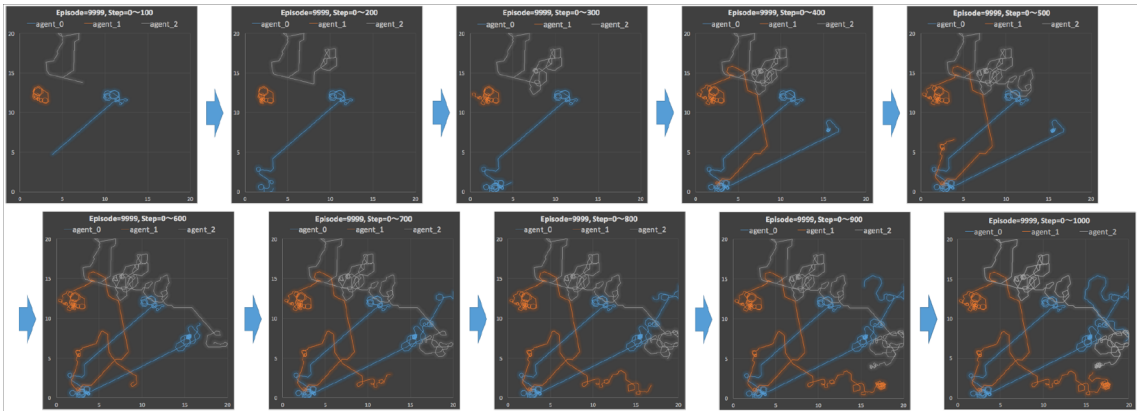

図 12：RQL エージェントが学習を通して獲得した壁や他個体との衝突を回避するための同じ場所での回転行動の例（100 ステップ毎の分解図）。

この結果は、各個体が独立に概念を形成し、かつ明示的なフィードバックが行えない原始社会的状況において、暗示的フィードバックがコミュニケーション・システムの形成に対して優位に貢献する重要な要素の 1 つになり得るということを示唆するものである。

4 つの研究によって、原始的コミュニケーションの創発メカニズムを明らかにすることを試み、個体の学習能力としては、過去の基礎的行動履歴、およびその順序を学習・予測する強化学習能力が重要な役割を果たし得ること、またコミュニケーションの成立に重要なその他の要素としては、暗示的フィードバックが寄与し得る可能性があることを明らかにした。これらの研究成果は、原始的なコミュニケーションの創発についての議論を今後更に深めることに貢献するものであると考えられる。しかし、原始的コミュニケーション創発メカニズムの解明にはまだ至っておらず、これから更なる研究が必要である。

５．主な発表論文等

〔雑誌論文〕（計 3 件）

(1) 佐藤 尚、基礎的行動プリミティブの履歴情報学習によるコミュニケーションの萌芽、計測と制御、査読有、53 巻 9 号、2014、pp. 847-852

(2) Takashi Sato、Emergence of Proto-Communication using Action Primitives Symbolized in Recurrent Q-Learning Agents, Journal of Information and Communication Engineering, 査読有, Vol. 2(2), 2016, pp. 87-93

(3) Takashi Sato, A Comparative Study on the Performances of Q-learning and Neural Q-learning Agents toward Analysis of Emergence of Communication, Journal of Information and Communication Engineering, 査読有, Vol. 2(4), 2016, pp. 128-135

〔学会発表〕（計 5 件）

(1) 佐藤 尚, 基礎的行動プリミティブの履歴情報学習によるコミュニケーションの萌芽, 計測自動制御学会システム・情報部門学術講演会 2014, 2014 年 11 月 21～23 日, 岡山大学津島キャンパス（岡山県・岡山市）

(2) 白﨑 史子, 佐藤 尚, 衝突回避ゲームにおけるコミュニケーション創発現象の解析のための Q 学習および Neural-Q 学習エージェントの性能比較, 計測自動制御学会システム・情報部門学術講演会 2014, 2014 年 11 月 21～23 日, 岡山大学津島キャンパス（岡山県・岡山市）

(3) 佐藤 尚, 基礎的行動強化学習に基づくコミュニケーション創発現象の解析のためのエージェントモデルの構築, 複雑系科学×応用哲学 第 2 回沖縄研究会, 2015 年 8 月 19～21 日, 琉球大学（沖縄県・中頭郡西原町：19, 21 日）, 沖縄工業高等専門学校（沖縄県・名護市：20 日）

(4) Takashi Sato, Symbolization of action primitives in recurrent-Q learning agents playing a collision avoidance game, The 21st International Symposium on Artificial Life and Robotics, 2016 年 1 月 20～22 日, B-Con PLAZA（大分県・別府市）

(5) 麓 有喜, 佐藤 尚, 暗示的フィードバックに基づくコミュニケーション成功率の向上, 一般社団法人電子情報通信学会ニューロコンピューティング研究会, 2017 年 6 月 23～25 日, 沖縄科学技術大学院大学（沖縄県・国頭郡恩納村）

６．研究組織

(1) 研究代表者
佐藤　尚（SATO, Takashi）
沖縄工業高等専門学校・
メディア情報工学科・准教授
研究者番号：70426576

(2) 研究協力者
橋本　敬（HASHIMOTO, Takashi）
北陸先端科学技術大学院大学・
知識科学系知識マネジメント領域・教授
研究者番号：90313709