博士論文

自律ロボットのための
情動に基づく意思決定システムに関する研究

Study on a Decision Making System
Based on Emotions for Autonomous Robots

平成 28 年 3 月

綿　田　　将　悟

山口大学大学院理工学研究科

目次

概要

より高度で人間らしいロボットシステムの開発のために生物がもつ情動に焦点を当て、情動のモデル化やその工学的応用に関する研究が数多く存在する。生物の情動は、情動が持ち主の内部状態や外部環境を柔軟で複合的に表現できる点や、他者とのコミュニケーションを円滑にしている点で、種々の工学分野で応用可能であり、魅力的な研究分野である。人工情動に関する研究の多くは人間が進化の過程で既に獲得した情動機能をモデル化している。しかし、人間が情動発生のメカニズムを進化の過程や後天的な経験で環境に適応して獲得したように、自律的なロボットの開発の観点においては、環境に適応的な人工情動の構成方法が必要であると考えられる。そこで、本研究では、人間の情動反応を模倣させるのではなく、必要な情動反応をロボット自身が外部環境に適応して形成することが可能なシステムとして、マルコフ情動モデルに基づくロボットの意思決定システムを提案した。

提案システムは、ロボットシステムに入力された感覚刺激を認識し、それに応じて情動状態の更新（情動反応）を行い、情動状態により動機づけされる行動（情動行動）を決定し、その情動行動を反映した制御信号を出力する、これらの一連の処理の流れを実装する枠組みである。一般的にこれらの処理を実装する上では、認識された感覚刺激と情動反応の対応付け、および、情動反応と情動行動の対応付けの各ルールが記述される必要がある。しかし、提案システムにおいては、これらのルールを人間の情動に基づいて事前に手動設計する従来の人工情動システムと異なり、ロボットに与えられたタスクやその動作環境に適応的に自動設計する機能を有する。提案システムは次に述べる 2 つの情動学習過程により情動反応および情動行動に関するルールを構築する。1 つ目の情動形成学習は感覚刺激と情動反応の対応付けを行う学習過程である。この学習は自己組織化写像（SOM）を用いた感覚刺激のクラスタリングによって行われる。2 つ目の情動行動学習は情動反応と情動行動を関連付ける学習過程である。この学習はタスクの試行におけるシステムの設計パラメータを適切に調節することにより行われる。システムに関する評価シミュレーションの結果では、2 つの情動学習過程により適切なロボットの行動決定法が自動構築され、有効な行動決定が生成されることが確認された。しかし、システムにおいて情動形成学習と情動行動学習の 2 つのプロセスが独立して行われることにより、感覚刺激の学習の際に、情動行動を行った際にロボットが得られる感覚刺激の発生確率が考慮されていない。そのため、タスクの実行時に発生頻度が低い感覚刺激に対しても情動反応への対応付けが行われ、行動決定法の性能が低下する可能性がある。

そこで、次に、発生頻度の低い感覚刺激への対応付けを減少させることを目的に、ロボットの経験に基づく情動の再形成学習を導入した改良型システムを提案した。以前のシステムでは情動形成学習は感覚刺激の予測値に基づく学習サンプルを用いて情動行動学習と独立に行われていたが、改良型システムでは、情動行動学習の際にロボットがタスク中に得た感覚刺激をオンラインで同時に学習可能となるようにシステムの部分的な改良が行われた。改良型システムの評価シミュレーションの結果より、情動の再形成学習によりタスク中の発生頻度が低い感覚刺激への不必要な対応付けが減少し、生成されるロボットの行動決定法の性能を向上させることが示された。

ここまでの研究では、システムにより設計される行動決定法は学習する行動タスクに依存するため、目的とする行動タスク毎に学習する必要があった。システムにおける学習過程は膨大な繰り返し計算が必要であるため、この問題点は実機ロボットにおける学習を困難とする場合があると考えられる。そこで、最後に、提案システムのさらなる汎用性を示すために、提案システムの強化学習への応用を行った。システムに基本的な行動学習のための強化学習を導入し、情動は学習を効率的に行うための補助的な役割をもつ学習戦略を提供する、新たな枠組みを提案した。ここで、学習戦略とはタスクに依存せず、あらゆるタスクに再利用可能な知識の効果的な利用方法を指す。本研究ではシステムにおける学習戦略を強化学習におけるメタパラメータの適応的制御とした。応用システムの評価シミュレーションの結果より、適切なメタパラメータの制御器がシステムにより自動的に設計され、獲得された制御器を用いることで強化学習手法による行動学習を効率化することが確認された。さらに、提案システムは従来のメタパラメータ制御手法よりも複雑な制御ルールを表現可能であることが示された。

本研究の成果として、情動反応と情動行動のルールを自動的に生成可能なシステムおよびその学習手法を提案した。さらに、強化学習への応用システムの提案により、実際のロボットへの対応におけるシステムのより汎用性の高い利用方法を示した。これらの成果は、ロボットのより自律的な動作の実現に貢献すると考えられる。

第1章　　緒言

1.1 はじめに

近年、人口減少や高齢化等により、医療や福祉の様々な分野における人的作業の補助を務める社会貢献型の高度な機能を有するロボットの開発が進み、注目を浴びている[1]。更には、産業用ロボット、癒し型エンターテインメントロボット等の開発も盛んである。このようにロボットがヒトの生活に深く関わるようになるにつれ、それを制御するシステムにはこれまでに重要視されてきた正確かつ高速な処理を行う能力以外に、ヒトとの円滑なコミュニケーションを行う能力や、環境への適応能力が強く求められるようになった[2][3]。これらの背景により、ロボットの制御システムはヒトによりプログラミングされた制御から次第に自律的な動作を備えたものへと進化してきている。即ち、ロボットが自身の感覚センサーの入力に対して行う行動に関して、あらかじめ組み込まれた行動様式に従って単に行動を行うのではなく、ロボット自身の経験によりその行動様式の学習・推論することが求められる。

このような自律的な動作を実現する手段の一つとして、最も知的な情報処理装置、即ち、人間の脳機能、特に、学習機能･記憶機能･適応機能をモデル化し、これらを利用した自律型脳型計算機の開発が盛んに行われている。脳のモデル化に関する研究の歴史は古く、1943年に発表された W.S. McCulloch と W. Pitts の神経細胞モデルに始まり、伊藤正男の小脳モデル（1958 年）、F. Rosenblatt のパターン識別器であるパーセプトロン（1962 年）などの数多くの数理モデルが発表された。しかし、当時の計算機の能力による手法の非実用性から工学的に魅力的でないとされ、研究熱が下火となった時期もあった。また、ロボットの制御システムにはオンラインで情報を処理する能力が重要となるが、一般的に学習、進化、適応のアルゴリズムを適用する場合、膨大な繰り返し計算が必要となり、オンライン処理が困難となる。しかし、計算機の計算能力の飛躍的向上と、インターネットの普及による学習データ収集の効率化により、近年、脳型計算機に関する研究は再び隆盛期を迎えることとなる。さらに、通信技術がより身近になった現代では、ロボットのオンライン性に関する問題への対策の一つとして、無線通信機能を利用するリモートブレイン方式も用いられる。ロボットの各種センサーから得た入力を無線通信で高性能な制御計算機へ送り、その制御計算機で繰り返し計算を行い得られた行動様式を無線通信でロボットに送り返すことでオンライン処理が可能になる。

一方で、これらの脳をモデル化した脳型計算機に関する研究分野では、生物が持つ情動の概念に焦点をあてた研究が盛んに行われている。生物の情動は、情動が持ち主の内部状態や外部環境に影響を受けて柔軟に表現することができる点、他者とのコミュニケーションを円滑にしている点で、種々の工学分野で応用可能であり、魅力的な研究分野である。生物における情動の魅力的な特徴として、まず、情動が動機付けの働きを持ち、行動決定に大きく影響している点を挙げる。例えば、人は、体温が低い（内部状況）時、冷たい水（外部状況）に触れると、不快（情動）に感じ、冷たい水を避ける（適切な行動）、といったように、内部や外部の状態を複合的に処理し情動として表現し、それに対する適切な行動を誘発する働きを持っている。この機能をロボットへ応用することにより、制御システムはより複雑な行動様式を記述でき、その行動ルールをヒトが理解しやすい形で表現することが可能となる。次に挙げる特徴は、コミュニケーションにおける役割である。人は他者との情報伝達の際に、喜怒哀楽のような情動を表現することにより、より多くの情報を伝えることが出来る。ロボットがヒトの職場や生活空間に進出する場合、ヒトとロボットがより円滑にコミュニケーションを行うためのヒューマン・エージェント・インタラクション（HAI）が重要となる。これらの理由から、情動の概念をロボットシステムへ導入することの有用性が期待されている。

また、人間の情動発生のメカニズムは進化の過程で獲得されたものであり、同時に、後天的な経験に基づいて適応的に変化すると考えられる。例えば、人間が高所に上ると恐怖を感じるのは、進化の過程で危険に対する防御反応として情動発生のメカニズムが備わったと考えられる。また、明るい家庭で育った子はよく笑うのは、後天的な経験により情動生成のメカニズムが変化していると考えられる。情動のロボットへの応用に関しても、情動生成のメカニズムは適応的に形成されることが自律ロボット開発の観点では望ましい。最近では、情動を模擬した機能を搭載したロボット製品が数多く登場している。しかしながら、これらのロボットのほとんどが、予めプログラミングされたルールに従い、画一的な情動反応を表現したものである。このように、自律システムの開発の視点から見ると、人工情動に関する研究はまだまだ未発展だといえ、今後研究を進めていく必要があると考えられる。

1.2 研究背景

近年の人工知能やロボット工学の分野における技術発展は著しく、これまでモノとしてヒトに使用されてきたロボットは将来的に、自身で学習し、考え、自らの意志で行動する、まるでヒトのパートナーのような存在となると予想されている。その際に求められるロボ

ットの自律性や人間らしさの実現には、ロボットへの人工情動の実装が重要であると考えられる。ロボットの人工情動に関する研究は大きく分けて「感情認識」、「感情表現」、「情動生成」の３つに分けられる。ここで「感情」と「情動」の２つの言葉が用いられるが、情動は比較的短時間の感情の動きを意味し、情動は「感覚刺激に伴う反応」、感情は「反応の出力結果」と考えられる。

感情認識と感情表現の 2 つの研究テーマは、ヒトとロボット間の双方向情報交換の実現において核となる技術であり、世界的にその重要性が認識されている。ヒトの表情と感情との関係性については、心理学の分野において古くから研究されており、ロボット工学における感情認識と感情表現の研究は心理学における知見に基づいたものが多い。感情推定の代表的な研究としては、音声や顔画像、脳波などの入力データに特徴抽出手法や識別器を用いることでヒトの感情を推定する研究が挙げられる。手法としては動画像のオプティカルフローに注目した研究[4]や、識別器にニューラルネットワークを用いたものがある[5][6]。また、感情表現の研究では、普遍性のある表現の法則を見出すために、人間に与える印象の要因と印象の内容との関係を明らかにする研究が行われている。これらの研究は、ヒト型ロボットもしくはスクリーンに映し出された人工的な顔画像に人間のような演出を提供し、システムとヒトとのインタラクションを円滑にする効果が示されている[7]。また、感情表現の関連研究として、ヒトがロボットから受ける心理的影響を調査する研究も行われている[8]。

情動生成の研究においては、ロボットの外部からの刺激に対する情動的評価とそれに基づくロボットの情動的行動を実現するため、人間の情動をモデル化した様々な情動モデルが提案されている。これらの研究は脳神経学観点の研究と心理学観点の研究に大別される。脳神経学観点の研究においては、Moren らは刺激に対して情動評価を行う計算的学習モデルとして、視野、感覚野、眼窩前頭皮質、扁桃体の４つの脳部位モデルから構成される情動モデルを提案した[9]。その内の１つである扁桃体モデルは様々な分野で応用が試みられており、滑車の角度と位置の制御[10] [11]や電熱マイクロ熱交換器の制御[12]、連想記憶システム[13]、時系列予測[14]、報酬系に適用した強化学習システム[15][16]に扁桃体モデルが組み込まれ、いずれも良い結果を示している。一方、心理学的観点からは、次のような研究がおこなわれている。2002 年に Michaud らによりロボットの意図的な行動選択のための情動による動機づけシステム EMIB[17]が提案されたことが端緒となり、その後さまざまな人工情動モデルが提案された。EMIB は基本的な行動生成を担う Behavioral レベル、センサー入力に対する反応や過去の記憶に基づいた行動選択を行う Recommendation レベル、そして行動の持続性を提供する Motivation レベルの３層のモジュール群により構成される複雑な

構造を持つ。また、Banik らはマルコフ過程に基づく確率的手法として情動モデルを提案し、その情動状態をグループタスクにおける各ロボットの役割分担に利用した[18][19]。Banik らの手法ではロボットの行動決定への利用方法が明確に定義されてはいないが、システムは簡潔で汎用性が高い。同時期に Qing-mei らも有限オートマトンを用いた情動モデルを提案している[20]。また、Daglarli らは隠れマルコフモデルに基づく情動モデルを提案した[21]。このモデルでは、情動状態は過去の行動状態を用いて更新され、情動状態は行動の動機付けを定義するものとして、動作の継続時間の決定に使用された。このように、脳における情動は人間の長期的な意思決定計画のための重要な役割を果たしていると考えられている[22]。Christopher らは刺激・応答のパターンを高速に処理する即応性の情動（reactive emotions）と、長期の行動計画に影響を与える熟慮的な情動（deliberative emotions）のハイブリッド構造の情動モデルを提案した[23]。また、Xue-fei らはニューラルネットワークの強化学習機構に基づいた人工情動モデルを提案した[24]。これは強化信号（報酬もしくは罰）として情動状態を用いることで、別々の個性を持つロボットが障害物回避実験によって様々な学習結果を得ることができた。同様に、Ahn も個体の個性に焦点を当てた情動モデルを提案し、仮想ロボットシステムにおいて個性に基づく様々な情動表現が可能であることを示した[25]。また、Zhang らはヒトとのインタラクションによる漸次的な情動学習手法を提案した[26]。ヒトが自然画像を見た時の脳波と画像特徴を用いてファジィニューラルネットワークの一種である ANFIS（Adaptive neuro-fuzzy inference system）により学習し、ヒトが自然画像を見た時の複雑な情動反応をシステムが自動的に模倣することを示した。このように人工情動に関する研究は様々な視点から情動をモデル化し、多様なタスクに応用されている。

一方、本研究室においても 2005 年から現在に至るまで、自律的なロボットシステム開発のための人工情動に関する研究を進めている。過去の研究では、Moren らが提案した扁桃体モデル[9]をベースとし、その出力を情動反応としてマルチカオスニューラルネットワークのネットワーク制御に用い、情動反応による動的記憶システムの想起能力の向上を示した[27] [28]。また、扁桃体モデルを強化学習法に応用することで、より生物の脳構造に近い情動の概念を導入した強化学習を提案した[15] [16]。これは、扁桃体のモデルのパラメータを変更することにより生物ごとに異なる感情の個性を持たせることができ、ジレンマが存在する学習タスクにおいてパラメータを変えることにより様々な行動パターンを獲得することを示した。さらにその応用として、人間の視覚から得られる色彩情報と情動との関係性を用いた強化学習の報酬決定法への拡張が行われた[29]。また、情動の概念を利用した移動ロボットの経路決定に関する研究を行い、マルチエージェントタスクにおけるロボット間の協調制御への拡張を行った[30][31][32]。

1.3 研究の目的

これまでに人工情動の生成に関する多くの先行研究について述べたが、これらは主に、人間のような柔軟で不確定要素のある行動決定や表現の実現や、情動反応の実装によるロボット動作の最適化、そして、人間の情動反応を模倣することによる複雑な情動反応の解明を目的としている。それらの研究のほとんどは人間がもつ情動の機能を分析・モデル化したものである。我々人間をはじめ、生物はそれらの情動発生のメカニズムを進化の過程や後天的な経験で環境に適応して獲得したと考えられている。例えば、人間の「高いところに立つと恐怖を感じる」という情動反応は進化の過程により得られたが、もし仮に地球に重力が発生しなければ、そのような情動反応は不必要であり、獲得されなかった可能性がある。このように人間が種の保存に有用な情動反応を適応的に獲得したように、自律的なロボットの開発の観点においては人工情動の生成においても環境に適応的なシステムの構成方法が必要であると考えられる。そこで、本研究では、人間の情動反応を模倣させるのではなく、必要な情動反応をロボット自身が外部環境に適応して形成することが可能なシステムとして、マルコフ情動モデルに基づく自律ロボットの意思決定システムを提案する。

提案システムは、システムに入力された感覚刺激を認識し、それに応じて情動状態の更新（情動反応）を行い、情動状態により動機づけされる行動（情動行動）を決定し、その影響を受けたロボットの動作を実現する制御信号を出力する、これらの一連の処理の流れを実装する枠組みである。一般的にこれらの処理を実装する上では、認識された感覚刺激と情動反応の対応付け、および情動反応と情動行動の対応付けの各ルールが記述される必要がある。しかし、提案システムにおいては、これらのルールを人間の情動に基づいて事前に手動設計するのではなく、システムを実装するロボットのスペックや対象とするタスクに適応して自動設計する機能を有する。システムは次に述べる 2 つの情動学習過程により情動反応および情動行動に関するルールを構築する。1 つ目の情動形成学習は感覚刺激と情動反応の対応付けを行う学習過程である。この学習は自己組織化写像（SOM）を用いた感覚刺激のクラスタリングによって行われる。2 つ目の情動行動学習は情動反応と情動行動を関連付ける学習過程である。この学習はタスクの試行におけるシステムの設計パラメータを適切に調節することにより行われる。

次に、提案システムの改良のためにロボットの経験に基づく情動の再形成学習を導入する。感覚刺激の予測値に基づく学習サンプルのみならず、ロボットがタスク中に得た感覚刺激をオンラインで学習可能となるようにシステムの部分的な改良を行う。再形成学習に

より感覚刺激と情動反応の対応付けをより適切に修正することで、生成されるロボットの行動決定法の性能向上を目的とする。

最後に、提案システムのさらなる汎用性を示すため、システムの強化学習への応用を行う。システムに基本的な行動学習として強化学習を導入し、情動は学習を効率的に行うための補助的な役割をもつ学習戦略を提供する新たなシステムを提案する。学習戦略とは、学習時における行動方策やメタパラメータの決定のように、タスクに依存せず、あらゆるタスクに再利用可能な知識の効果的な利用方法を指す。従って、応用システムの提案により、実際のロボットへの対応におけるシステムのより汎用性の高い利用方法を示す。

1.4 論文の構成

本論文の構成を以下に紹介する。各章の関連図は図.1.1 で表される。第 1 章の緒言では研究背景と目的について述べる。第 2 章では情動生成とその工学的応用に関する従来研究について説明する。第 3 章では提案するマルコフ情動モデルに基づく意思決定システムついて説明する。また、マルチロボットによる環境同定問題に関するシミュレーションによるシステムの性能評価について述べる。ここでは提案システムにおける情動反応の振る舞いや、タスクの学習によりロボットの行動決定法が自動的に獲得されことを示す。第 4 章ではロボットの経験に基づく情動反応の再構成に関して述べる。提案システムを情動反応の再構成に対応するように修正し、移動ロボットの実験によりロボットの状況と情動反応の対応付けが自身の経験に基づき更新されることを確認する。第 5 章では提案システムの強化学習手法への応用に関する研究成果について述べる。最後に第 6 章では結論として本研究の総評と将来展望を述べる。

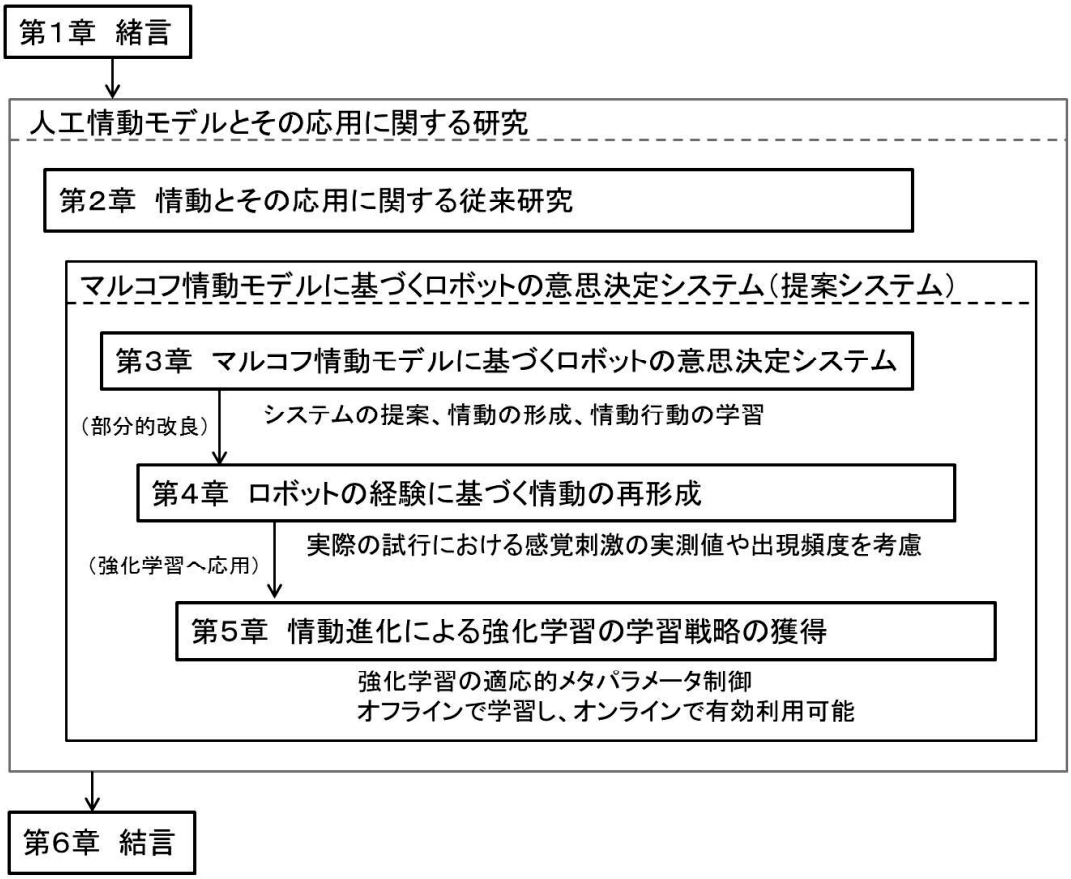

第1章　緒言
人工情動モデルとその応用に関する研究
第2章　情動とその応用に関する従来研究
マルコフ情動モデルに基づくロボットの意思決定システム（提案システム）
第3章　マルコフ情動モデルに基づくロボットの意思決定システム
（部分的改良）
システムの提案、情動の形成、情動行動の学習
第4章　ロボットの経験に基づく情動の再形成
（強化学習へ応用）
実際の試行における感覚刺激の実測値や出現頻度を考慮
第5章　情動進化による強化学習の学習戦略の獲得
強化学習の適応的メタパラメータ制御
オフラインで学習し、オンラインで有効利用可能
第6章　結言

図 1.1: 各章関連図

第2章　　情動とその応用に関する従来研究

本章では、本研究におけるテーマである情動について、その概念と生体における働き、そして情動の工学分野への応用に関する先行研究について述べる。まず、2.1 節では情動について大域的な定義や人間における情動の働きについてのいくつかの説を述べる。次に、2.2 節では心理学の分野における情動のモデル化に関する研究について述べ、最後に、2.3 節では工学分野における人工感情モデルとその応用に関するいくつかの研究を紹介する

2.1 情動とその働き

情動という言葉は本来心理学用語であり、その定義には異論が多い。しかし、集約すると、情動は個体及び種族維持に関する感情体験と身体反応を意味する[33]。人間の脳の中には外界から加えられた感覚刺激を、その時の体内環境情報と比較して、その個体及び種族保存の観点からの重要性を決定する情報処理が行われる。この情報処理および生体反応を情動と呼び、反応の出力結果を感情と呼ぶ。また、脳神経学においては、大脳辺縁系で本能的な情である情動（エモーション・喜怒哀楽）を生じ、これが上位の大脳新皮質によって昇華されて感情（フィーリング）となって創出されるとしている[34]。

Rolls[35]は生体における情動の役割を具体的に整理し、次のように提唱した。

① 情動は行動への動機づけを発生させる
情動によって誘起された動機づけ信号が行動決定機構への入力されることにより、正の強化刺激に対してはそれを獲得しようとして働き、負の強化刺激ならばそれを回避しようとする。情動により動機づけられる行動を情動行動と言う。

② 情動状態は行動を介して個体の生存確率を高める
個体にとって快い感じを起こす刺激（空腹時の食物、低体温時の高温刺激）は正の強化刺激であり、それを獲得することは個体の生存率を高めることになる。同様に、不快感を起こす刺激は負の強化刺激であり、それを回避することで生存率を高める。

③ 情動状態は柔軟性のある行動を可能にする。
ある強化刺激に対して生ずる情動状態は、いくつかの行動反応の中からその状況下で最も適切な行動反応を起こすことができる。

④ 情動状態は行動遂行のための身体の準備をする
情動状態は自律および内分泌反応を誘発することにより、情動行動を円滑に行えるような身体の準備状態を作る。

⑤ 情動表出は個体間の「通信手段」の役割を果たす。

威嚇行動や怒りの表出は相手に警告を与え、不必要な闘争の発生を抑える。このように、情動表出反応は生物にとって進化的な安定を目指す適応戦術の一つである。

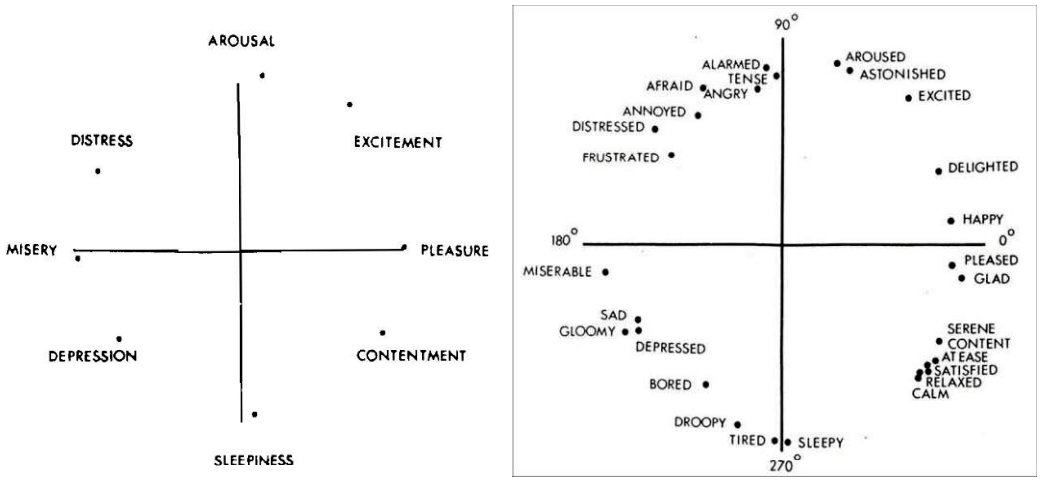
⑥ 情動状態は種族の生存確率を高める。

典型的な例として、子が両親に、親が子に、そして親同士がそれぞれ抱く愛着の情動である。生物の群居生活や社会生活も、その群や種族の遺伝子の生存確率を高める適応手段とみなすことができ、これらの行動は情動の関与なしには発現しない。

2.2 心理学分野における情動のモデル化に関する研究

心理学的観点における情動の研究は、脳内メカニズムを考慮せず情動を単に心理的状態として捉え分析する。いずれも情動は感情の最小単位であるという考えに基づき（基本情動という）、複数の基本情動から様々な感情状態を表現することができる。ここで使われる「感情」という言葉は「情動」と似た意味をもつが、特に喜怒哀楽のような明確に発言する強い感情状態を主に対象とする。

Rusell は「快」と「活性」の二つの基本感情を用いて全ての感情状態を扱うことができるとしている。Russell は、すべての感情は「快（Pleasure）－不快（Unpleasure）」、「活性（Arousing）－不活性（Sleepiness）」の 2 次元平面上に、円環上に並んでいるとする円環感情モデルを提唱した[37]。Rusell の円環感情モデルを図 2.1 に示す。図 2.1(a)は基本情動により構成される 2 次元平面を表しており、(b)は同じ平面に感情状態を配置した図である。相関の高い似た感情状態は近くに配置されており、Happy や Pleased など、喜びに関連する感情状態が Pleasure の基本情動の近くにあることが分かる。



(a) 基本構成　　(b) 表現される 28 の感情状態

図 2.1: Russell の円環感情モデル

また、Rusell の円環感情モデル以外にも、様々な基本情動を元に構築した感情モデルが提案されている[38]-[49]。心理学の分野で提案された感情モデルの一覧を表 2.1 に示す。いずれも基本情動をブレンドまたはミックスすることにより、喜びや悲しみなどの感情（非基本的な情動）を表現している。しかし、モデルの構造は同じであっても根源となる思想は異なる。例えば、Plutchik は進化の道筋をたどっていくと顔の表情は次第に少なくなっていくが、逆に、他の身体システムを使った情動の表現は多く残っていると指摘し、進化の時間軸を導入した立体モデルを提案した[41]。Panksepp はネズミの脳の部位を電気刺激する実験に基づき、基本的情動反応パターンであるパニック、怒り、期待、恐れを観察した[42]。Johnson-Laird と Otaley はヒトが情動について話すときに使う言葉を調べるというアプローチをとった[46]。

このように、様々な感情・情動モデルが提案されていることは、感情を明確なモデル化することが容易でないことを意味している。これらの心理学モデルを工学的に用いることは有用であり、その際には感情の表現能力も重要であるが、各軸にどのような変数を割り当てるか、軸の数（次元数）をいかに減少して多様な感情を表現できるか、などが重要要素となる。

表 2.1: 心理学の分野で提案された情動モデル

提案者	年代	基本情動
McDougall	1908	「怒り」「嫌悪」「得意」「恐れ」「服従」「愛情・慈愛」「驚き」の7つ。
Watson	1930	「恐れ」「愛情」「怒り」の3つ。
Arnold	1960	「怒り」「忌避」「勇気」「落胆」「欲望」「絶望」「希望」「恐れ」「愛情」「悲しみ」「憎悪」の11つ。
Plutchik	1980	「受容」「予期」「喜び」「恐れ」「怒り」「悲しみ」「驚き」「嫌悪」の8つ。
Panksepp	1982	「期待」「恐れ」「怒り」「パニック」の4つ。
Tomkins	1984	「興味」「苦痛」「怒り」「喜び」「軽蔑」「恥」「驚き」「嫌悪」の9つ。
Rolls	1986	「喜び」「怒り」「恐れ」「安らぎ」の4つ。
Weiner＆Graham	1984	「喜び」「苦しみ」の2つ。
Johnson-Laird＆Oatley	1987	「怒り」「嫌悪」「不安」「喜び」「悲しみ」の5つ。
Gray	1991	「不安」「肯定的情動」「怒り・恐怖」の3つ。
Izard	1991	「怒り」「軽蔑」「嫌悪」「苦痛」「恐れ」「罪」「興味」「喜び」「恥」「驚き」の10つ。
Ekman	1992	「怒り」「嫌悪」「恐れ」「喜び」「悲しみ」「驚き」の6つ。

2.3 工学分野における人工感情モデルに関する研究

工学分野においても、脳神経学や心理学の知見を基に工学的応用のための人工情動モデルが提案されている。その中でも、ロボットの行動決定手法への応用を目的とした、情動生成に関わる先行研究をいくつか紹介する。

2.3.1 扁桃体における情動学習のモデル化

Moren らは扁桃体の入出力の機能をモデル化した扁桃体モデルを提案した[9]。扁桃体モデルは視床、感覚野、眼窩前頭皮質、扁桃体の 4 つの脳部位から構成され、感覚刺激の受容から感覚刺激の価値評価までの流れを表現したものである（図 2.2）。扁桃体モデルは感覚入力を受け取り、視床（Thalamus）および感覚野（Sensory Cortex）を通じて眼窩前頭皮質（Orbitofrontal Cortex）と扁桃体（Amygdala）へと送られる。扁桃体の反応は感覚入力に対するパターン想起によって決まり、その出力である情動値はパターンを評価する役割を果たす。また、Obayashi らはこの扁桃体モデルと Russell の円環感情モデルを組み合わせたシステムを提案した。環境モデルから得られる各情報を用いて、円環感情モデルで設定された基本感情の数 N 個存在する扁桃体モデルからそれぞれ情動値を導き出し、円環感情モデルが持つ感情マップから現在の感情状態を導き出す。出力された感情状態をエージェントの行動決定に利用することで、生物自身が持つ感情状態を意思決定の基準に加えた複雑な意思決定が可能となった。

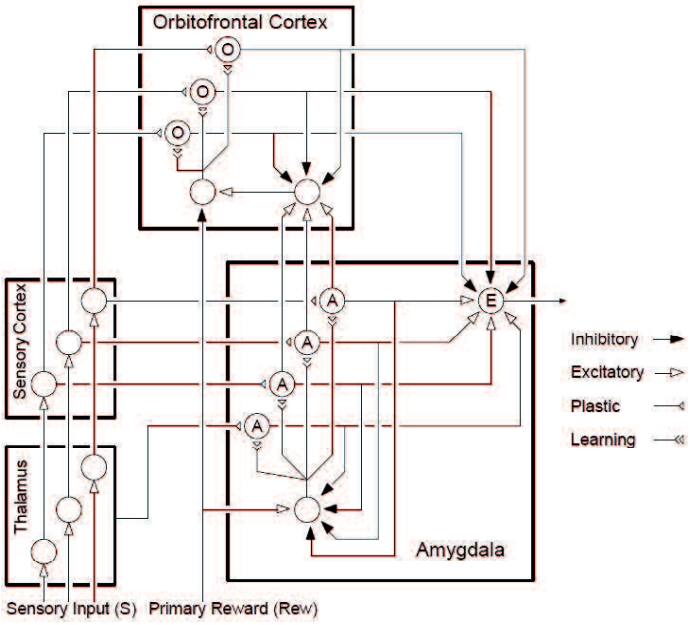


図 2.2: Moren の扁桃体モデル
（文献[9]より引用）

2.3.2 移動ロボットのグループタスクにおける情動に基づく役割分担

Sajal Chandra Banik らはマルコフ過程に基づく確率的手法として情動モデルを提案した[18][19]。提案されたシステムにおける情動発生器では、入力刺激 u を非線形変換することで情動誘発因子 $\alpha, \beta, \gamma, \delta$ を生成し、マルコフ情動モデルによって情動値 Y を更新する（図2.3）。このシステムはマルチロボットによるごみ回収タスクのシミュレーションにおいてテストされた。結果では、一体のロボットが故障した際に他のロボットが故障したロボットの役割をフォローするといった、マルコフ情動モデルに起因する各ロボットの効果的な役割分担が行われ、マルコフ情動モデルの有用性が示されている。

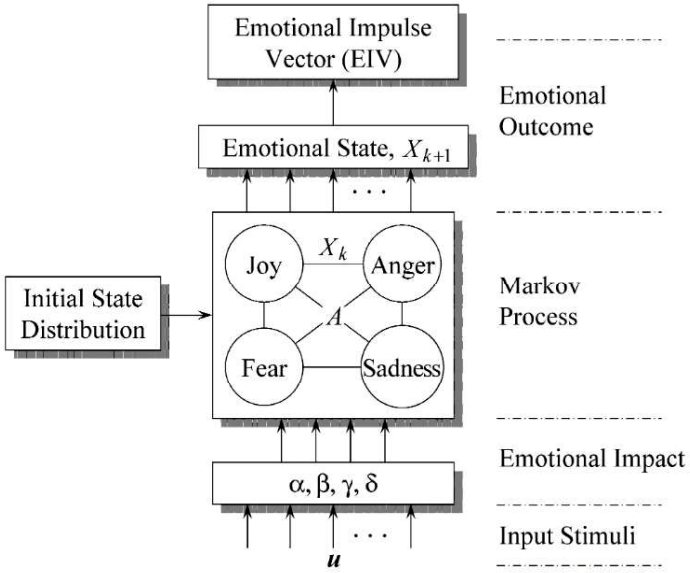


図 2.3: Banik の情動状態生成システム
（文献[18]より引用）

2.3.3 人工情動を用いた移動ロボットのナビゲーションモジュール

Christopher らは刺激・応答のパターンを高速に処理する即応性の情動（reactive emotions）と、長期の行動計画に影響を与える熟慮的な情動（deliberative emotions）のハイブリッド構造の情動モデルを提案した[23]。即応性と熟慮性のハイブリッド構造をもつ行動学習システム[50]は以前にも提案されているが、このモデルは人間において情動が移動行動へ及ぼす影響に関する知見に基づき、ロボットの移動計画に特化して設計された手法である。即応性情動はロボットの移動速度の制御と障害物回避反応の抑制に影響を与え、具体的には、ロボットが障害物に近づくほど即応性の恐怖情動が上昇し、減速の行動が誘発されることで障害物との衝突リスクを下げる。熟慮的情動は経路決定や探索計画に影響を与え、具体的

には、ロボットの衝突が発生した領域に関して熟慮的の恐怖情動が上昇し、そのような危険領域を回避する経路決定を促す。

2.3.4 強化学習におけるメタパラメータ制御と神経修飾物質系との関係の仮説

銅谷らは強化学習システムのメタパラメータと脳内物質である神経修飾物質とを関係づける仮説として強化学習における TD 誤差、割引率、逆温度定数、学習係数の 4 種類の指数やメタパラメータが脳内のドーパミン系、セロトニン系、ノルアドレナリン系、アセチルコリン系の神経修飾物質と対応付けられているという理論を提示した[51] [52] [53]。また、脳科学の分野においては、それらの神経修飾物質は情動の発生に関与していることも解明されている[36]。銅谷の一説に基づき、水野らは神経修飾物質系に対応付けた強化学習のメタパラメータ制御法を提案した[54]。報酬の減少によりメタパラメータを制御することで環境の急激な変化に追従できることを示した。一方、水野らは TD 誤差を用いてメタパラメータを調整する学習法を提案した[55][56]。メタパラメータ制御に関する従来手法の事前に設定すべきパラメータが多いという問題点に対して、設定が必要なパラメータを一つに減らすことにより、強化学習法への適応を容易にしている。また、秋口らは報酬と罰の最大化及び最小化に関する複数の Q 値を持つ目標選択型 Q-Learning を提案し、状況に応じて学習目標を選択的に変更することによって複雑な情動行動を学習するシステムを実現した[57]。

第3章　　マルコフ情動モデルに基づく意思決定システム

本章では、提案するマルコフ情動モデルに基づく意思決定システムについて述べる。まず、3.1 節で提案システムの特徴と狙いについて述べる。次に、3.2 節では提案システムの全体構造と各モジュールの役割について述べる。3.3 節では提案手法における行動選択法及び情動反応に関するルールの自動設計法について述べる。3.4 節では提案手法の評価のために、マルチロボットによる環境同定問題に関するシミュレーションについて述べる。最後に 3.5 節では考察を述べる。

3.1 システムの特徴と狙い

従来の人工情動の生成に関する研究では、人間が進化の過程で既に獲得した情動の機能をモデル化したものである。つまり、システムにより認識された感覚刺激と情動反応の対応付けや、情動反応と情動行動の対応付けの各ルールを人間の情動に基づき事前に設計する必要があった。しかし、自律ロボット開発の観点ではロボットにおいても外部環境に適応して情動を獲得することが望ましいと考えられる。そこで、本研究では、人間の情動反応を模倣させるのではなく、必要な情動反応をロボット自身が外部環境に適応して獲得することを目的とした、ロボットの情動に基づく意思決定システムを提案する。

提案システムは、システムに入力された感覚刺激を認識し、それに応じて情動状態の更新（情動反応）を行い、情動状態により動機づけされる行動を決定し（情動行動）、その影響を受けたロボットの動作を実現する制御信号を出力する、これらの一連の処理の流れを実装する枠組みである。提案システムにおいては、これらの処理に必要な対応付けルールを人間の情動に基づいて事前に手動設計するのではなく、システムを実装するロボットのスペックや対象とするタスクに適応して自動設計する機能を有する。システムは次の 2 つの情動学習過程によりそれらのルールを設計する。1 つ目の情動形成学習は感覚刺激と情動反応の対応付けを行う学習過程である。2 つ目の情動行動学習は情動反応と情動行動を関連付ける学習過程である。これらの学習過程により情動反応及び情動行動に関するルールを自動的に設定することで、設計者はロボットの行動決定法をタスクやロボットに応じて手動設計する必要がなく、ロボットは環境に適応的に行動決定法を自動獲得すると考えられる。また、獲得された行動決定法は人間の情動反応と同様な表現で意味的に説明・理解することが可能である。

3.2 システムの構成

本節では提案システムの全体構造と、各モジュールの役割について述べる。まず 3.2.1 節ではシステムの全体構造とタスク中における処理の流れについて述べ、その後、3.2.2 節で Emotion モジュールについて、3.2.3 節では Emotional Behavior モジュールについてその詳細を説明する。

3.2.1 システムの全体構成と処理の流れ

提案するマルコフ情動モデルに基づく意思決定システムは Cognition、Emotion、Emotional Behabior、Robot Controller の 4 つのモジュールにより構成される。システムの全体の構成図を図 3.1 に示す。ここで、Y は情動状態を表す情動ベクトルであり、式(3.1)のように 4 つの基本情動の情動値により構成される。Z は感覚刺激であり、ロボットのセンサーやカメラからの入力情報を意味する。$\alpha, \beta, \gamma, \delta$ は情動誘発因子であり、情動モデルにおける各基本情動を誘発する働きをもつパラメータである。例えば、α は Joy の基本情動を誘発する。X は行動選択確率ベクトルであり、ロボットがとり得る各行動の選択確率により構成される。s はロボットの内部状態を表し、u はロボットへの制御出力を表す。

$$Y = \begin{bmatrix} y_{joy} & y_{anger} & y_{fear} & y_{sadness} \end{bmatrix}^T \tag{3.1}$$

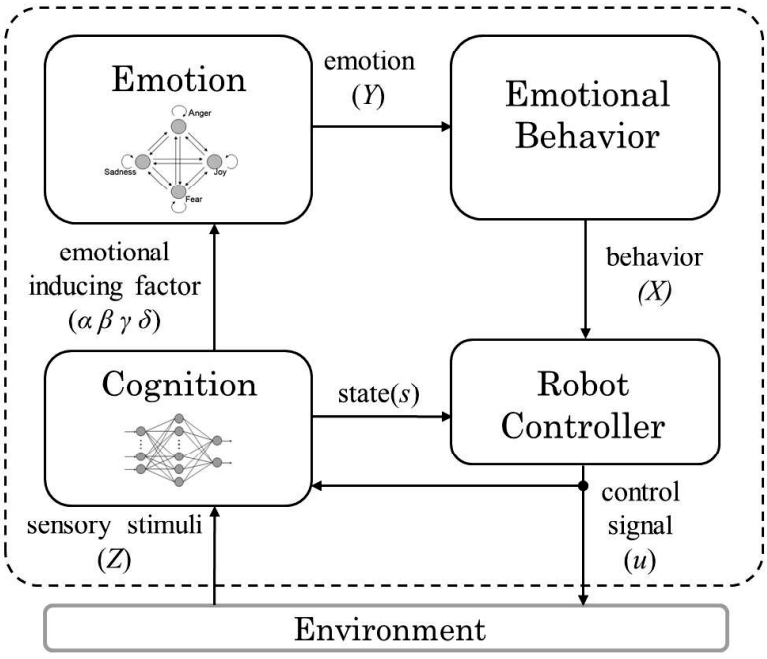


図 3.1: マルコフ情動モデルに基づく意思決定システム（全体図）

４つの各モジュールの機能は、以下の通りである。

Cognition モジュール

外部環境からの感覚刺激 Z やロボットの内部状態 s を認識し、情動誘発因子 $\alpha, \beta, \gamma, \delta$ を決定する。

Emotion モジュール

Banik[18]らによって提案されたマルコフ情動モデルによって構成される。Cognition モジュールが生成した情動誘発因子 $\alpha, \beta, \gamma, \delta$ を用いて情動状態確率 Y を更新し、Behavior Selection モジュールへ出力する。

Emotional Behavior モジュール

情動状態確率 Y を利用して行動選択確率 X を決定する。

Robot Controller モジュール

ロボットの内部状態 s を考慮して、行動選択確率 X に基づき決定された行動を実現するロボットの制御入力 u を出力する。例えば、「立ち上がる」という行動を実現するために、現在のロボットの姿勢に応じた、モータへの出力トルクの計算を行う。

タスク実行時のシステム全体の処理の流れは次のようになる。処理のフローチャートを図 3.2 に示す。ロボットが環境から受け取った感覚刺激は Cognition モジュールにより情動誘発因子（$\alpha, \beta, \gamma, \delta$）に変換される。次に、情動誘発因子に基づき Emotion モジュールの情動状態を表す情動ベクトル Y が変化する。このように感覚刺激によって情動状態が変化する処理を情動反応と言う。また、この情動反応により特定の基本情動 e の情動値 y_e が上昇することを、e の情動が誘発されると言う。次に、Emotional Behavior モジュールでは情動ベクトル Y を用いて行動選択確率ベクトルを求め、それに基づき選択された行動を Robot Controller モジュールに出力する、もしくは、行動選択ベクトルを Robot Controller モジュールに出力する。このように情動状態 Y を用いて決定された行動、つまり、情動に動機づけられた行動を情動行動という。Robot Controller モジュールは行動選択確率ベクトル X に基づき、ロボットが情動行動を実行する制御出力を生成する。ロボットのタスク実行時には、これらの処理を繰り返し実行することで、環境に応じて逐次的に行動を決定する。

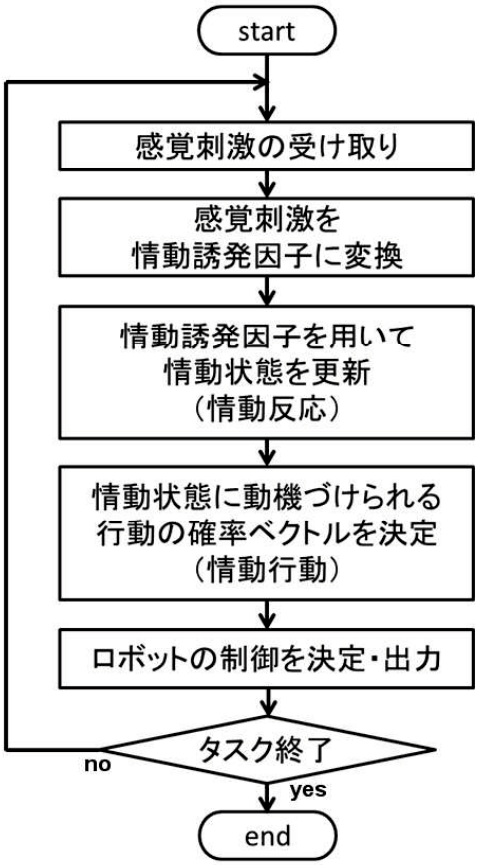


図 3.2: タスク実行時のシステムの処理のフローチャート

次に 3.2.2 節では Emotion モジュールについて詳細を説明する。

3.2.2 EMOTION モジュール

Emotion モジュールは、Banik[18] [19]らによって提案されたマルコフ情動モデルによって構成されている。Banik らによって提案されたマルコフ情動モデルを搭載した情動生成システムは 2.3.2 で説明した。

マルコフ情動モデルは、心理学の観点に基づく、複数の状態を示すオートマトンから構成される確率モデルであり、４つの基本情動（Joy、Anger、Fear、Sadness）から成り立っている。マルコフ情動モデルの構造を図 3.3 に示す。

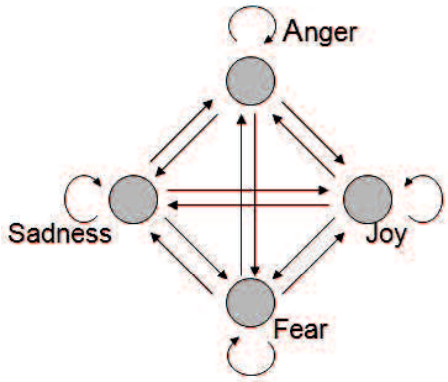


図 3.3: マルコフ情動モデル

情動ベクトルの各成分は、４つの基本情動と対応しており。

このモデルの情動ベクトル Y の遷移は式(3.2)で行われる。

$$Y_{k+1} = CY_k \tag{3.2}$$

C は情動遷移行列であり、式(3.3)で表される。

$$C = \begin{bmatrix} P_{joy/\,joy} & P_{joy/\,anger} & P_{joy/\,fear} & P_{joy/\,sadness} \\ P_{anger/\,joy} & P_{anger/\,anger} & P_{anger/\,fear} & P_{anger/\,sadness} \\ P_{fear/\,joy} & P_{fear/\,anger} & P_{fear/\,fear} & P_{fear/\,sadness} \\ P_{sadness/\,joy} & P_{sadness/\,anger} & P_{sadness/\,fear} & P_{sadness/\,sadness} \end{bmatrix} \tag{3.3}$$

ただし、行列 C の各要素は状態間の遷移確率であり、例えば $P_{S/R}$ は状態 R から状態 S への遷移確率である。これらの要素 P は感覚刺激の影響を受けてオンラインで変化する。

例として、*Joy* からの遷移確率 $P_{*/joy}$ の決定方法を式(3.4)、式(3.5)、式(3.6)、式(3.7)に示す。

$$P_{anger/joy} = q_{anger/joy} + (\beta - \alpha)q_{anger/joy} \quad (3.4)$$

$$P_{fear/joy} = q_{fear/joy} + (\gamma - \alpha)q_{fear/joy} \quad (3.5)$$

$$P_{sadness/joy} = q_{sadness/joy} + (\delta - \alpha)q_{sadness/joy} \quad (3.6)$$

$$P_{joy/joy} = 1.0 - (P_{anger/joy} + P_{fear/joy} + P_{sadneess/joy}) \quad (3.7)$$

ここで、$\alpha, \beta, \gamma, \delta$ は情動誘発因子であり、基本的な情動状態 Joy, Anger, Fear, Sadness をそれぞれ喚起する役割を持ち（例えば α は *Joy* を喚起するパラメータである）、感覚刺激や内部状態を反映して Cognition モジュールにより決定される。q は情動遷移の事前確率行列である Q の要素であり、事前に決定されるべき値である。遷移確率行列 Q を式(3.8)に示す。

$$Q = \begin{bmatrix} q_{joy/joy} & q_{joy/anger} & q_{joy/fear} & q_{joy/sadness} \\ q_{anger/joy} & q_{anger/anger} & q_{anger/fear} & q_{anger/sadness} \\ q_{fear/joy} & q_{fear/anger} & q_{fear/fear} & q_{fear/sadness} \\ q_{sadness/joy} & q_{sadness/anger} & q_{sadness/fear} & q_{sadness/sadness} \end{bmatrix} \quad (3.8)$$

以上の過程により、情動確率ベクトル Y は情動誘発因子 $\alpha, \beta, \gamma, \delta$ により、内部状態を反映して更新される。

3.3 情動学習によるルールの自動設計法

3.2 節では、ロボットが環境から受け取った感覚刺激に対して情動反応および情動行動の決定が行われる処理の枠組みについて説明した。また、タスク実行時におけるこれらの処理の流れについて述べた。しかし、これらの処理を実行する上では、Cognition モジュールにおける認識された感覚刺激と情動反応の対応付け、および、Emotional Behavior モジュールにおける情動反応と情動行動の対応付けの各ルールをタスク実行前に記述する必要がある。本研究の目的はこれらの記述を自動化することであり、ロボットが学習により環境に適応してこれらのルールを獲得することが望ましい。そこで、本論文では情動形成学習と情動行動学習の２つの学習プロセスを提案する。

提案システムにおいては情動形成学習と情動行動学習の２つのプロセスによって行動決定法が設計される。情動形成学習は「どのような状況の時、どのような情動が誘発されるか」を決めるものであり、Cognition モジュールにおける感覚刺激と情動誘発因子の対応付けのルールを学習する。情動行動学習は「どのような情動状態の時、どのような行動が選択されやすいか」を決めるものであり、情動状態とそれに動機づけされる情動行動の対応付けのルールを学習する。

次節から、２つの学習プロセスを実現するための Cognition モジュールおよび Emotional Behavior モジュールの実装方法を提案する。Cognition モジュール

情動形成学習は Cognition モジュールにおける SOM を用いた感覚刺激のクラスタリングによって行われる。Cognition モジュールは多層パーセプトロン（Multilayer Perceptron：以下 MLP）によって構成される（図 3.4）。MLP は生物の脳のニューロンをモデル化した人工ニューラルネットワークの一種であり、Rumelhart らの誤差逆伝播法の学習アルゴリズムにより、入力パターンに対する出力パターンの教師有り学習が可能である[58]。近似能力や汎化能力に優れ、数多くの応用例が存在する。

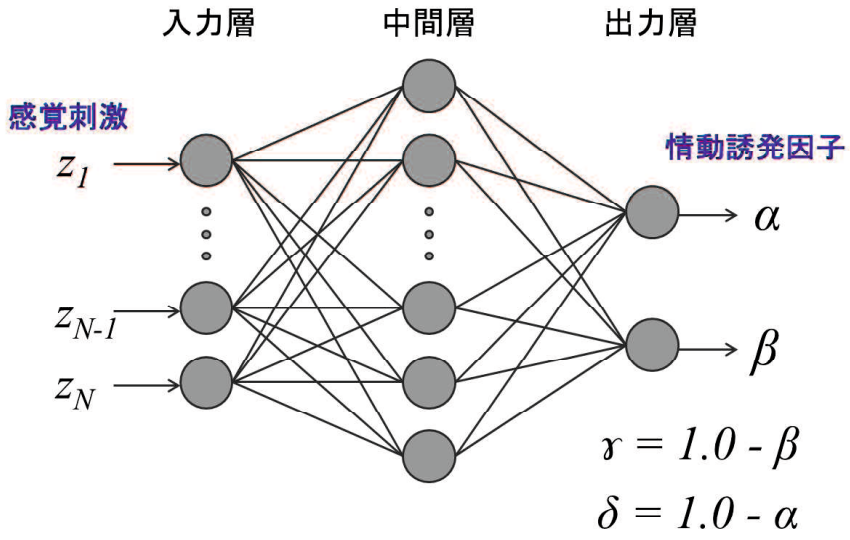


図 3.4: Cognitive モジュールの構造

Cognition モジュールでは MLP の入力層へロボットの感覚刺激を入力すると、MLP の順伝搬処理により出力層から情動反応を起こす情動誘発因子が出力される。この時、これらの変換法は MLP の学習能力により教師データを用いて自動的に設計される。Cognition モジュールを構成する MLP は出力層が 2 つのノードにより成り、それぞれ Joy を喚起する α と

Anger を喚起する β に対応する。Sadness、Fear を喚起するパラメータ γ、δ はそれぞれ対立する情動 Joy、Anger を喚起するパラメータの逆確率で求められる。入力層の数は感覚刺激の最大数である n、中間層の数はタスクによって調整可能であるとする。

MLP は高い近似能力と汎化能力を持つが、学習には適切な教師データが必要である。そこで、MLP の学習における教師データの生成に自己組織化マップ（Self-Organizing Maps : 以下 SOM）を使用する[59]。SOM は Kohonen により提案された人工ニューラルネットワークの一種であり、入力層と競合層の 2 層から成る。教師なし学習によって入力データを任意の次元へ写像可能であることが特徴である。

Cognition モジュールにおいて MLP の教師データの生成に SOM を使用する目的は、似た感覚刺激を似た情動誘発因子に対応するクラスタリングを実現することである。つまり、SOM の入力層へ入力した感覚刺激を、それらの性質の類似度に従って競合層にマッピングし、それを情動マップと見なすことでロボットが取り得る感覚刺激と情動反応の一意な対応付けを生成する。SOM を用いた MLP 学習の様子を図 3.5 に示す。

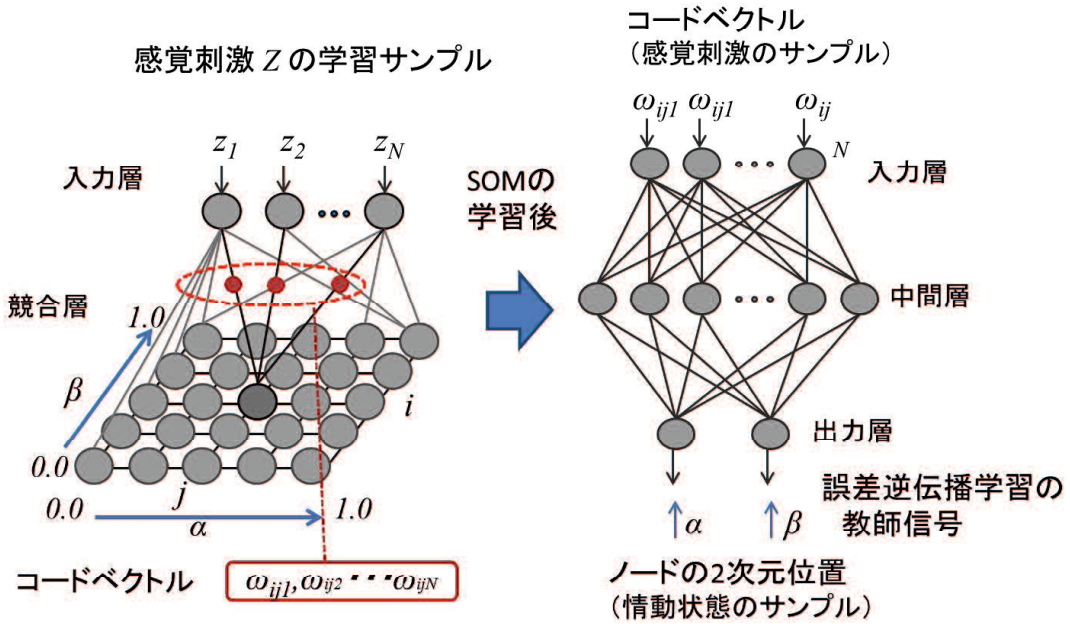


図 3.5: SOM を用いた MLP の学習プロセス

本研究では、SOM の競合層におけるノードの位置を情動誘発因子の値に対応させる。即ち、SOM の n 次元競合層（図 3.5 では $n = 2$）のノード位置を n 個の 0 から 1 の範囲に正規化された情動誘発因子（図 3.5 では α と β）で構成される n 次元空間と考えると、その感覚刺激

ベクトルに対する情動誘発因子の値が一意に決まり、学習後の n 次元 SOM の競合層各ノードのコードベクトル（結合荷重ベクトル）はそれぞれ類似した入力ベクトル群が圧縮されたものと考えることができる。入力層の n 番目のノードと競合層 i 行 j 列（k 番目）のノードの結合荷重の更新式を(3.9)式に示す。

$$\Delta\omega_{kn} = \eta h(k, k^*) \times (z_n - \omega_{kn}) \tag{3.9}$$

ここで入力ベクトルと距離が最小となるコードベクトルを持つ k*番目ノードを勝者ノードとする。$h(k,k^*)$は近傍関数であり(3.10)式に示す。ここで σ は分散であり学習の進行とともに小さくなる。よって学習により競合層の各コードベクトルは対応する入力ベクトルが圧縮された代表値へ収束していく。

$$h(k, k^*) = exp\left(\frac{-|k-k^*|^2}{\sigma^2}\right) \tag{3.10}$$

SOM の学習結果を MLP の教師データとして使用することについて、SOM の学習結果がコードベクトルの初期値に依存し、結果として競合層のノード位置座標である情動誘発因子値の不確実性が問題になるが、本提案システムでは MLP の学習後に、決定された情動誘発因子値をベースとした行動決定に関する確率行列 Q、A を GA により最適化する。よって感覚刺激と情動誘発因子の対応関係の不確実性の影響は GA の最適化の過程で吸収され、本システムにおける行動選択は決定された情動誘発因子をベースとした範囲で最適化されていると言える。また、本誌では視覚化のために 2 次元 SOM を使用し、そのため MLP の出力ノード数は 2 とした。しかし、SOM の多次元化により、全ての情動誘発因子と対になる 4 つの出力ノードをもつ MLP への拡張も可能である。

以上の過程により、感覚刺激の入力に対して、それらの状態空間を適切に反映した情動誘発因子 α, β を出力する MLP が構築される。本研究ではこの過程を情動の形成と呼ぶ。また、情動形成を終了した後のタスクにおける感覚刺激から情動誘発因子への変換は MLP のみを使用する。

3.3.1 EMOTIONAL BEHAVIOR モジュール

情動行動学習は Behavior Selection モジュールにおける行動決定法を定義するパラメータを GA により最適化することにより行われる。Behavior Selection モジュールにおける行動選択確率は情動ベクトル Y を用いて式(3.11)により決定される。ここで X は行動選択確率ベクトルで、例えば、ロボットが search、confirm、wait、return の 4 つの行動を取り得る場合には式(3.12)のように表され、ベクトルの各要素はロボットが取り得る行動の選択確率である。

$$X_{k+1} = AY_{k+1} \tag{3.11}$$

$$X = [x_{search} \quad x_{confirm} \quad x_{wait} \quad x_{return}]^{\mathrm{T}} \tag{3.12}$$

A は遷移行列で、上の例に従う場合、式(3.13)のように表される。

$$A = \begin{bmatrix} P_{search/\,joy} & P_{search/\,anger} & P_{search/\,fear} & P_{search/\,sadness} \\ P_{confirm/\,joy} & P_{confirm/\,anger} & P_{confirm/\,fear} & P_{confirm/\,sadness} \\ P_{wait/\,joy} & P_{wait/\,anger} & P_{wait/\,fear} & P_{wait/\,sadness} \\ P_{return/\,joy} & P_{return/\,anger} & P_{return/\,fear} & P_{return/\,sadness} \end{bmatrix} \tag{3.13}$$

行列の各要素は遷移確率であり、$P_{behavior\,/\,emotion}$ は情動状態 *emotion* から行動 *behavior* への遷移確率である。システムにおける情動反応と情動行動の対応付けは A のパラメータに依存し、A の設計により様々なロボットの行動決定法を表現できると考えられる。つまり、タスクを通して行動選択確率行列 A の各パラメータを適切に設定することができれば、タスクに対する適切な行動決定法を学習することが可能である。

本研究では行動選択行列 A の各パラメータの設定法として、近似会を探索するメタヒューリスティックアルゴリズムである遺伝的アルゴリズム（Genetic Algorithm：以下 GA）を使用する。GA は、データ（解の候補）を遺伝子で表現した「個体」を複数用意し、適応度の高い個体を優先的に選択して交叉・突然変異などの操作を繰り返しながら解を探索する進化計算手法である。A の各パラメータを GA の個体の遺伝子にコーディングし、そのパラメータを用いて実行したタスクの結果から個体の適応度を決定する。

3.3.1 システムの学習の流れ

提案システムにおける情動形成学習と情動行動学習の処理の流れは図 3.6 に示すフローチャートで表される。まず、情動形成学習では、まず、ロボットの感覚刺激を SOM によりクラスタリングすることにより、SOM の競合層へロボットの感覚刺激をマッピングする。この競合層に情動誘発因子を対応付けることで､競合層から MLP の学習サンプルを作成し、MLP の教師有り学習を行う。これらの過程により、感覚刺激から情動誘発因子への非線形変換法が自己組織化される。

次に情動行動学習では、GA による確率行列 A を最適化することにより、適切な行動決定法を獲得する。GA の各個体の適応度は実際のタスクの試行の結果に基づき決定される。

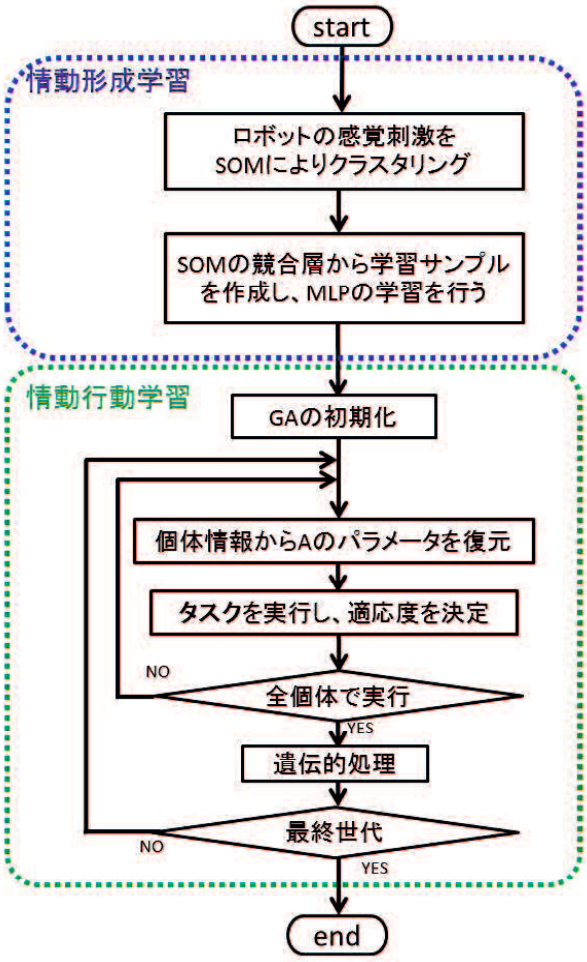


図 3.6: 情動学習のフローチャート

3.4 計算機シミュレーション

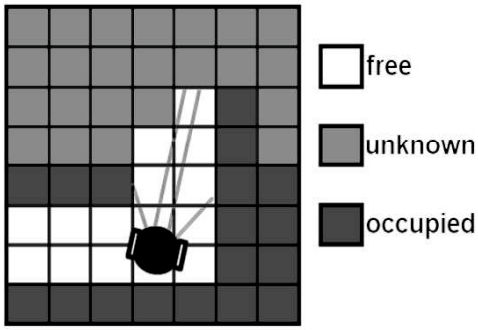
本章では、提案システムを用いたマルチロボットの行動計画シミュレーションについて述べる。まず、3.3.1 節では対象タスクとしたマルチロボットによる未知環境の環境同定問題について述べる。次に、3.3.2 節以降では目的の異なる 3 つの計算機シミュレーションについて述べる。

3.4.1 マルチロボットによる未知環境の環境同定問題

未知環境同定とは、未知環境内のすべてのエリアを探索し、正確な地図を生成するタスクである。一般的に、ロボットが環境内の現時点から目的地までの移動を実現するためには、経路決定や障害物回避のための手法が不可欠である[60]。ただし、これらの手法は環境の地図やロボットの自己位置（座標や方向などの姿勢）の情報が必要である。

環境が既知の場合、ロボットは事前に環境の地図を所有できる。しかし、構造が未知である環境においてロボットは移動しながら搭載しているセンサーやカメラと自己位置を用いて環境の地図を生成なければならない。さらに、ロボットが GPS やコンパスを搭載していない場合は、環境の地図生成と自己位置推定を同時に行う必要があり、このような問題は SLAM（simultaneous localization and mapping）と呼ばれる。災害現場のような構造が未知である環境において移動ロボットが作業を行う際には、地図の生成は重要なタスクであり、地図生成や SLAM に関する研究は現在においても盛んに行われている[61][62]。一方、臭気源探索ロボット[63][64]や警備ロボット[65]のような環境内を網羅的に移動する必要があるタスクにおいては、複数台のロボットグループにより構成されるマルチロボットを用いることが有効である。ただし、マルチロボットにおいては情報の共有方法や仕事の分担、交通の混雑などの更なる問題点が発生する[66]。未知環境同定問題にマルチロボットを適応する場合、特に地図の共有や探索地の適切な役割分担が重要となる。地図の共有においては、他ロボットから受け取った地図と自身の地図の結合を容易に行うために、地図のグラフ化や簡略化などの手法が数多く提案されている[67] [68]。また未知の環境を探索する場合、現在の探索状況に応じた適応的な役割分担が必要である[69]。

本シミュレーションのタスクであるマルチロボットによる未知環境の環境同定問題は、複数のロボットが迷路構造の環境内を探索し、到達可能なすべてのエリアに関する地図を作成するタスクである。ロボットは分散制御であり、各ロボット自身が地図を用いて行動決定を行う。また、ロボット同士は一定距離以内に位置するときに地図および位置情報の共有が可能である。本シミュレーションでは代表的な地図の種類である占有格子地図を想定する。占有格子地図は、環境を小さなグリッド（格子状）に分割し、グリッド毎に障害物の存在確率を保有するものである[60]（図 3.7 参照）。しかし、本シミュレーションでは複雑なタスクを対象とするために、各ロボットが正確なマッピングと自己位置推定が可能であることを前提とし、マッピング手法を簡略化した。即ち、環境は格子状のセルの集合で表現され、ロボットは同サイズのグリッドにより構成される占有格子地図を作成する。また、ロボットは環境上のセル間の状態遷移により移動するものとした。



(a) 各グリッドの状態(空間・未知・占有)

(b) 占有格子地図の例

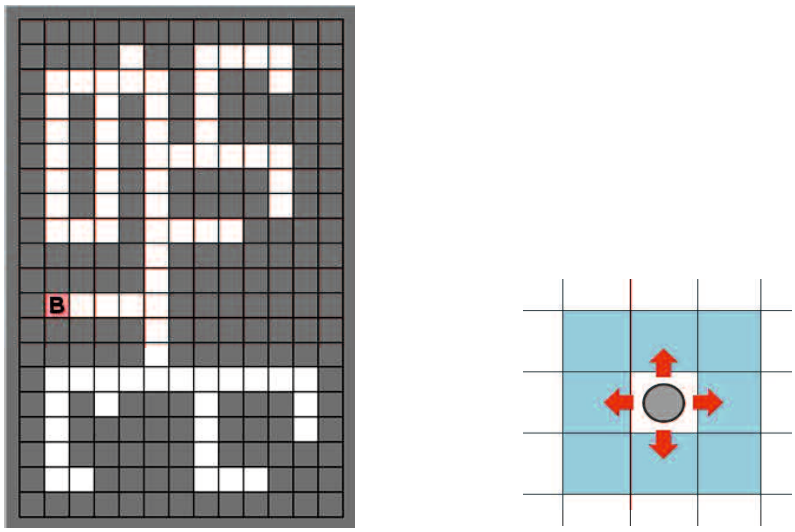
図 3.7: 占有格子地図

シミュレーションは目的の異なる 3 種類が行われた。まず、シミュレーション①では、提案システムの性質を調査するために情動形成学習と情動行動学習は行わず、情動反応と情動行動のルールを手動で設定し、タスクにおける情動反応や行動決定の振る舞いを観察した。次に、シミュレーション②では情動行動学習のみを行い、適切な行動を自動的に獲得できること確認する。最後に、シミュレーション③では情動形成学習と情動行動学習の両方を行い、感覚刺激に対する情動反応と行動決定が自動的に生成されることを確認する。

3.4.2 問題設定

対象とするタスクは、分散制御である 3 体のロボットで構成されるマルチロボットによる未知環境同定である。タスクの目的は複数のロボットがベースを出発し、環境内の全ての未知環境について地図生成を行い、その後、全ロボットがベースへ帰還することである。

シミュレーションは独自に開発したシミュレータによって行われた。シミュレータは Java 言語を用いて開発した。本シミュレーションでは、複雑なタスクを対象とするために、各ロボットが正確なマッピングと自己位置推定が可能であることを前提に、マッピング手法を簡略化した。即ち、環境及びロボットのローカルマップは図 3.8(a)のような格子状のセル集合で表現され、ロボットはセル間の状態遷移により移動する。セルの状態はロボットが走行できる通路（図中の白色）と走行できない壁（図中の灰色）、そしてロボットの出発地点であり、バッテリーの充電が可能であるベース(図中の B）が存在する。



(a) グリッド表現された環境　　(b) ロボットの移動範囲と認知範囲

図 3.8: シミュレーションにおける環境とロボットの表現

シミュレーションの開始後、各ロボットがベースから時間間隔をあけて出発し、環境内の全ての通路をマッピング後、全てのロボットがベースへ帰還することによりタスクが終了する。ロボットには、次の仮定をおく。

- 座標情報を持つが、方位情報は持たない。
- １ステップの行動で隣接する 4 近傍（上・下・左・右）のセルへ移動することができる。ただし、壁のセルへの移動は不可能である。
- センサーからの情報として、8 近傍のセルの状態を知覚しマッピング可能である（図 3.8-b）。
- 独自のローカルマップを作成しながら環境内を進み、一定距離（8 セル)内に存在する他のロボットとマップを共有することが可能。
- 環境内には 1 箇所ベースが存在し、ベースから出発し、ベースへ帰還する。
- ロボットはバッテリー値をもち、その初期値および最大値は 100 である。
- 移動することによりバッテリー値を消費し、バッテリー値が 0 になると、全ての行動が不可能になる。
- ロボットがベースに滞在している時、1 ステップでバッテリー値が 1 増加する。

ロボットはローカルマップを作成することによって環境についての知識を学習し、その地図からグラフ探索アルゴリズムを用いることにより、目的値への最短ルートを計算することが可能である。ロボットのローカルマップも環境と同様にセル集合で構成され、free（非占有）、occupied（占有）、unknown（未知）の 3 つのセル状態で表現される。タスク開始時点ではローカルマップの全セルが unknown に初期化されており、ロボットがセル状態を更新することをマッピングすると言う。またロボットはローカルマップ中の free セルに隣

接する unknown セルを frontier（未開拓：通路終端に隣接する未知セル）として認識し、行動や移動方向の決定に使用される。ロボットのローカルマップの様子を図 3.9 に示す。

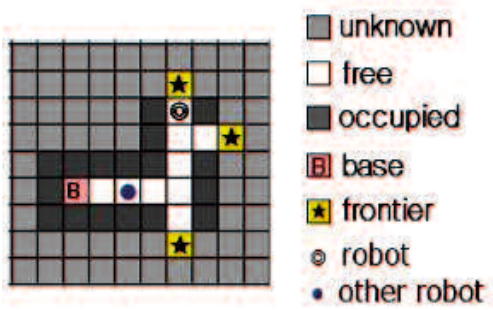


図 3.9: ロボットのローカルマップの様子

ロボットはそれぞれ次の 4 つの行動{ Search, Confirm, Return, Wait }のうち 1 つを選択し、その行動理念に基づいてエリアを選択し移動する。ここでエリアとは通路の分岐点で分けられる区分を定義し、木構造においては現在地を根とした場合の枝であり、それぞれのエリアに対して探索価値が計算される。探索価値はエリアへの距離とその近辺の frontier の数によって決定する。

Search

他のエージェントが目的地に設定していない、かつ最も探索価値の高いエリアを選択する。

Confirm

未マッピング地が多いエリアにおける探索の効率化のために、他ロボットにより既に目的地として設定しているエリアを選択する。

Return

ローカルマップに基づく base への最短経路を移動する。

Wait

移動せずに現在位置に待機する。

従って、行動選択確率ベクトル X は式（3.12）のように表される。また、Robot Controller モジュールは X を用いて最も選択確率が高い行動を選択し、適切な移動処理を出力する。

$$X = \left[x_{search} \quad x_{confirm} \quad x_{wait} \quad x_{return} \right]^T \qquad (3.12)$$

提案システムをマルチエージェントの環境同定問題に適応する場合、地図生成を行う Mapping モジュールの搭載が必要である。Mapping モジュールはロボットの視覚情報である近傍の座標の状態を入力し地図生成を行う。また、他ロボットとの地図の共有によって自身の地図を更新する。本シミュレーションにおけるロボットシステムの全体構造を図 3.10

に示す。また、Vは先天的な行動選択を促す本能ベクトルである。設計者によって事前に設定される行動決定法を意味する。本来、人間が欲求に任せて行動することが社会的評価を下げるように、情動だけに基づいた行動選択は好ましくない。例えば、食事中に満腹感を感じれば食事がまだ残っていても中断して別の行動をとることは好ましくない。同じく、ロボットにおいても、バッテリーの充電中に内部状態の変化により別の行動を選択しやすくなるが、実際には、充電を完了するまで待機するほうが効率的である。このような先天的知識を設計者与える場合にこのパラメータを使用する。

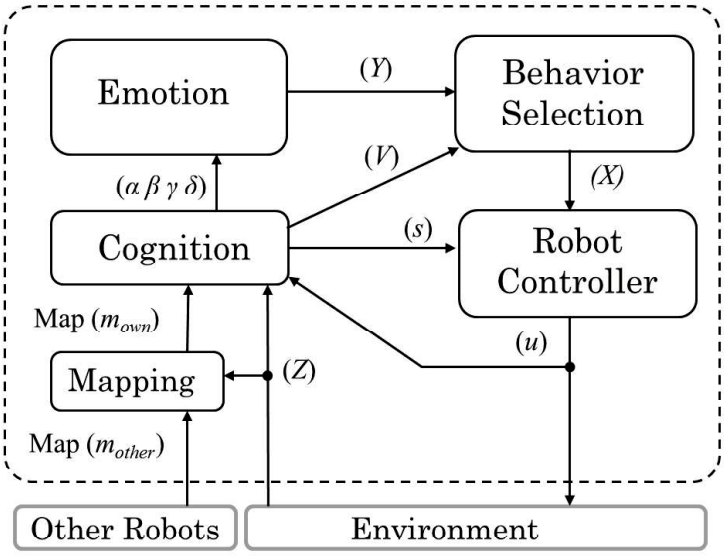


図 3.10: シミュレーションにおけるシステムの全体図

3.4.3 システムにおける情動反応と行動選択の観察のためのシミュレーション

この節で述べるシミュレーション①では、提案システムの性質を調査するために情動形成学習と情動行動学習は行わず、情動反応及び情動行動に関するルールを手動で設定したシステムを用いて、タスク実行における情動反応や情動行動決定の振る舞いを観察した。情動確率と行動状態の関係を明確にするために、Cognition モジュールにおける感覚刺激と情動反応の対応付け、及び確率行列 Q と A を設計者が自らのコンセプトに基づき手動で設定した。

感覚刺激 Z から情動誘発因子 $\alpha, \beta, \gamma, \delta$ への変換を次のように定義する。

Joy を喚起する α

自分のいるエリアの基地からの距離と、バッテリー残量によって決定する。バッテリーの残量が少なくなると値は減少する。また、基地から遠ざかるほど値は減少する。式(3.13)で定義される。

Fear を喚起する γ

ロボットが認知している未マッピング地の数によって決定する。未マッピング地の数が減ると上昇する。式(3.14)で定義される。

Anger を喚起する β

ロボットが認知している未マッピング地の数によって決定する。未マッピング地の数が増えると上昇する。式(3.15)で定義される。

Sadness を喚起する δ

基地からの距離と、バッテリーの残量によって決定される。バッテリーの残量が少なくなると値は増加する。また、基地から遠ざかるほど値は増加する。式(3.14)で定義される。

$$\alpha = (\text{(Battery) - (Distance from the Base)})/(\text{Max value of battery}) \quad (3.13)$$

$$\gamma = 1.0 * 0.3^{(Number\,of\,Frontein)} \quad (3.14)$$

$$\beta = 1.0 - \gamma \quad (3.15)$$

$$\delta = 1.0 - \alpha \quad (3.16)$$

遷移行列 A と Q の各要素である遷移確率は、設計者のコンセプトに基づいて試行錯誤によって決定された。遷移行列 Q、A、V のパラメータを式(3.17)に示す。

$$A = \begin{bmatrix} P_{search/Joy} & P_{search/anger} & P_{search/fear} & P_{search/sadness} \\ P_{confrimation/Joy} & P_{confrimation/anger} & P_{confrimation/fear} & P_{confrimation/sadness} \\ P_{return/Joy} & P_{return/anger} & P_{return/fear} & P_{return/sadness} \\ P_{wait/Joy} & P_{wait/anger} & P_{wait/fear} & P_{wait/sadness} \end{bmatrix}$$

$$= \begin{bmatrix} 0.5 & 0.8 & 0.0 & 0.0 \\ 0.5 & 0.1 & 0.8 & 0.1 \\ 0.0 & 0.1 & 0.1 & 0.8 \\ 0.0 & 0.0 & 0.1 & 0.1 \end{bmatrix}$$

$$Q = \begin{bmatrix} 0.5 & 0.1 & 0.2 & 0.1 \\ 0.1 & 0.7 & 0.1 & 0.15 \\ 0.25 & 0.1 & 0.5 & 0.15 \\ 0.15 & 0.1 & 0.2 & 0.6 \end{bmatrix}, \quad V_c = \begin{bmatrix} -0.2 \\ -0.2 \\ -0.2 \\ 0.6 \end{bmatrix}, \quad V_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad (3.17)$$

これらのパラメータの決定の際には、設計者のコンセプトが反映されている。例えば、A の設定に関して、Joy の情動は Search と Confirm を喚起するという概念に従って、$P_{search/joy}$と$P_{confirm/joy}$の確率を高く設定した。本能パラメータ Vに関しては、ベースにおいてバッテリーを充電している間のみV_c、それ以外ではV_0を出力するように設定した。最初に、図 3.11 に示す狭い環境を用いたシミュレーション結果について述べる。

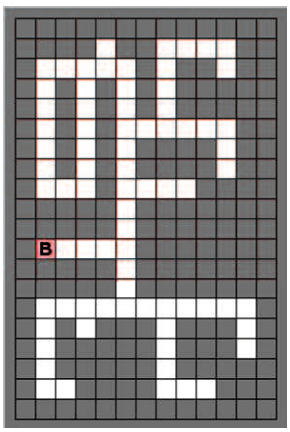


図 3.11: 狭い環境

各ステップ 1、5、9、13 におけるシミュレーションの様子を図 3.12 に示す。step1 は 1 台目、step 5 では 2 台目、step 9 では 3 台目のロボットがベースを出発した直後の様子である。step9 においては 2 代目のロボットが分岐点において、1 台目のロボットと違う方向である

下方向に曲がっている。同様に step13 では 3 台目のロボットが分岐点を上方向に曲がっていることがわかる。このようにシミュレーション中、其々のロボットは探索場所を分散しながら環境内のマッピングを行った。各ロボットが環境内のどの通路を訪問したのかを示すために、Fig.3.11 の通路セルを訪問したロボット別に色分けし、セル内にそのロボット番号を付加した図を Fig.3.13 に示す。図中の m は複数のロボットが訪問したことを示している。探索エリアを各ロボットが分担し、すべての通路を訪問していることがわかる。

シミュレーションにおける、一台のロボットの情動状態と情動行動の時間的変化のグラフを図 3.14 に示す。各情動の情動値と情動行動それぞれの変化を比較すると、情動行動の選択に情動状態が反映されていることがわかる。例えば、step45 周辺において Anger の情動値は下降、Fear の情動値は上昇し、それに伴って情動行動は Confirm に変化している。また、step45 におけるそのロボットのローカルマップの様子を図 3.15 に示す。凡例にある received は他のロボットから共有されたセル情報である。この時、ロボットが自身の近傍にある frontier をすべて解決しており、それにより Fear の情動が喚起され、他のロボットに割り振られたエリアを目標地とする Confirm へ行動を変更している。式(3.17)の状態遷移行列において、Fear からの状態遷移において、Confirm への遷移確率が高く設定されているため、結果より、設定したコンセプトに正しくシステムが機能していることがわかる。

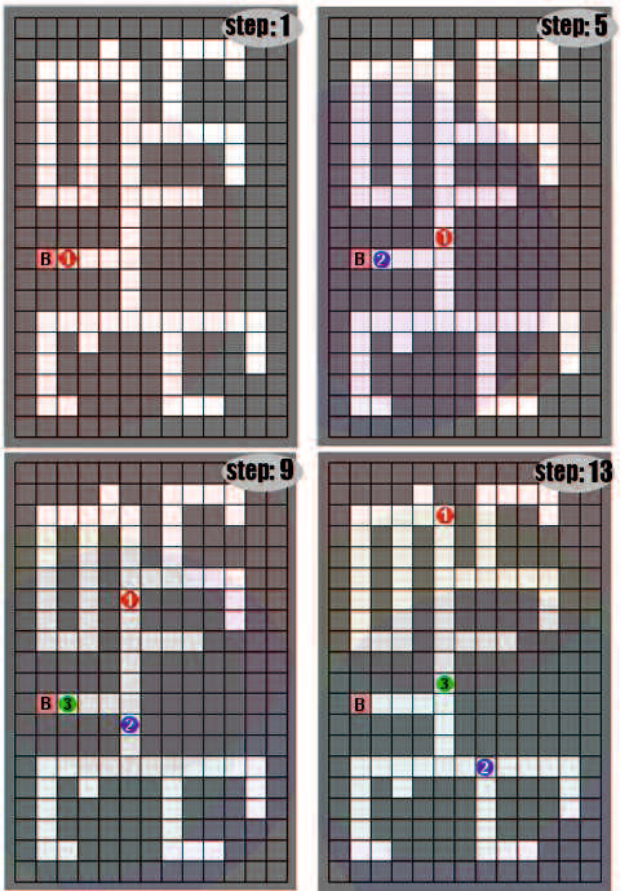


図 3.12: シミュレーションの様子

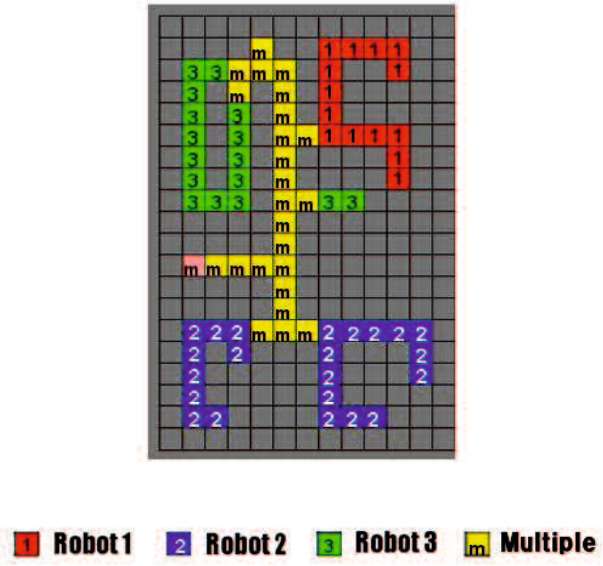
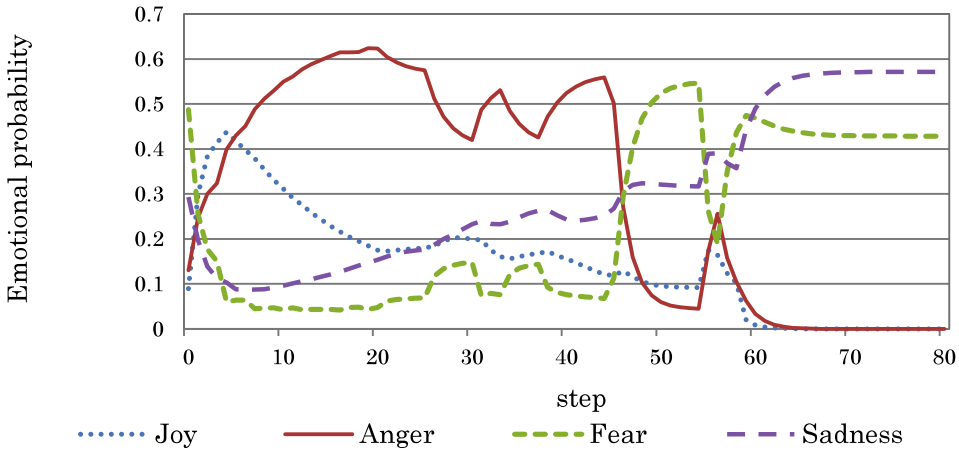
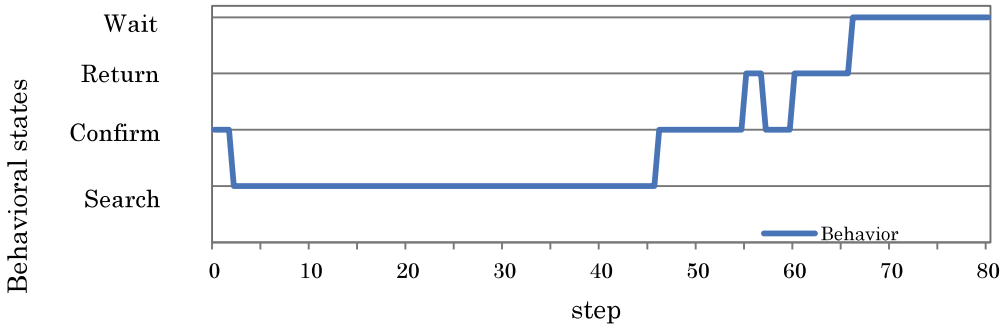


図 3.13: 各セルの訪問ロボットの番号



(a) 情動ベクトル値の推移



(b) 行動の推移

図 3.14: 1 ロボットの情動反応と選択行動の推移

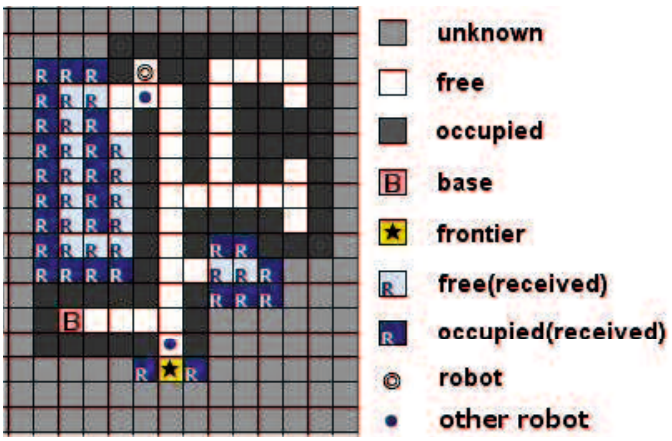


図 3.15: ステップ 45 のロボットのローカルマップの様子

次に、図 3.16 に示す広い環境を用いてシミュレーションを行った。この環境では各ロボットは探索中、一度以上ベースへ帰還し、バッテリーの充電を行わなければならない。環境内の各セルの訪問ロボットを表した地図を図 3.17 に示す。図中の m は複数のロボットが訪問したことを示している。各ロボットのみが訪問したセルの数はそれぞれ 33、34、33 であり、広い環境においても 3 体のロボットに平等に探索地が分担されていることが観察できる。シミュレーションにおける、一台のロボットの情動確率の時間的変化のグラフを図 3.18 に示す。

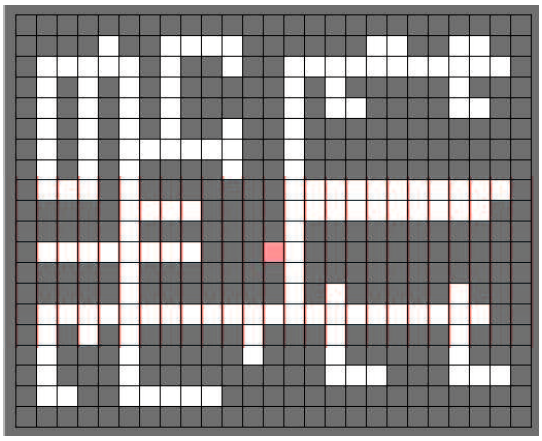


図 3.16: 広い環境

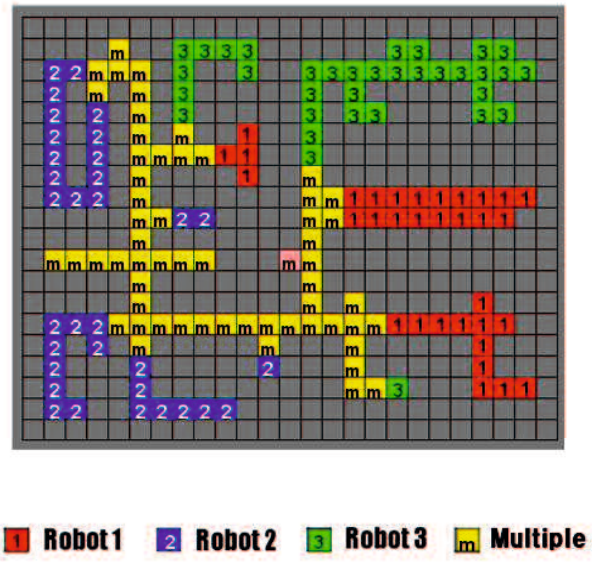


図 3.17: 各セルの訪問ロボットの番号

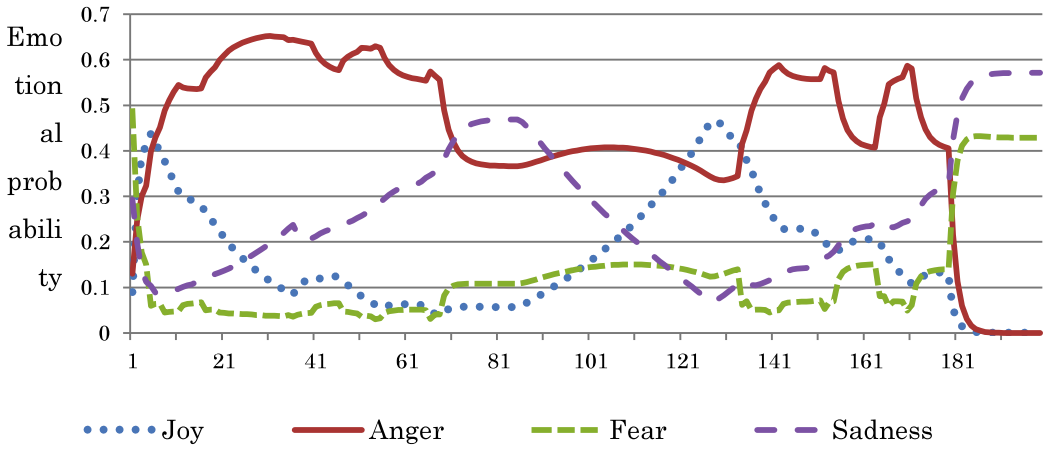
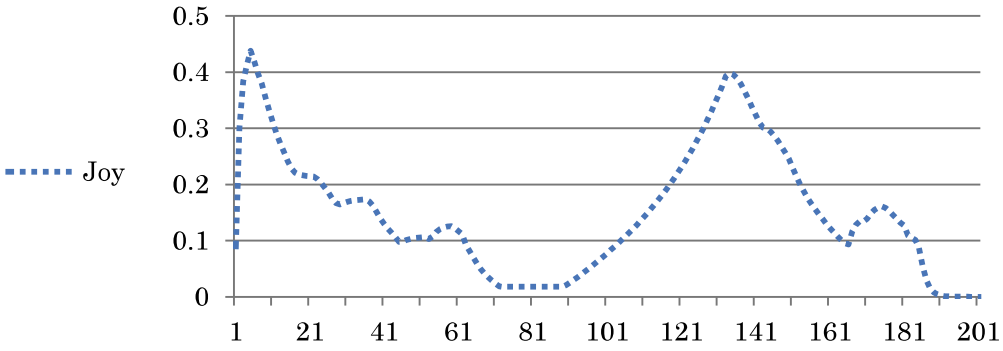
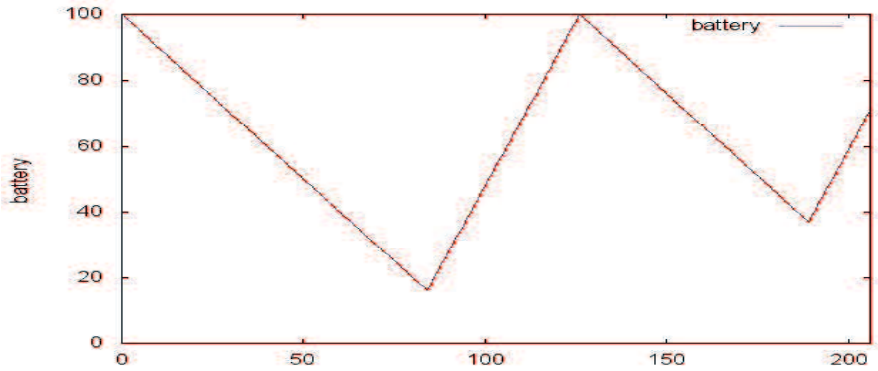


図 3.18: 1 ロボットの情動ベクトル値の推移

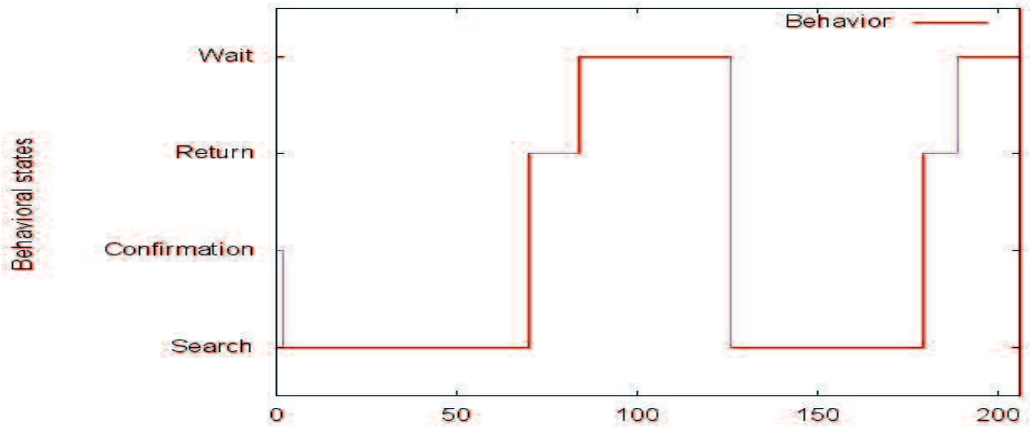
このシミュレーションで使用した環境が広いため、未マッピングエリアの数が多く、全体的に Anger の情動値が高いことがわかる。この中から Joy の情動に関して取り出し、Joy の情動値とバッテリー値、行動状態の各時系列変化のグラフを比較した。それぞれ 3 つのグラフを図 3.19 に示す。結果より、バッテリー値に関連のある情動値である Joy の状態確率は、バッテリーの値に対応して変化している。ロボットが行動を行っている時、バッテリーは減少し続け、基地に戻り Wait の行動（充電を行っている）時、バッテリー値は上昇し続けている。また、Wait はバッテリー値が最大値の 100 に達するまで選択されている。これは Cognitive モジュールの本能的行動を即発するパラメータである V によるものである。



a)　1 ロボットの Joy の情動値の推移



b)　1 ロボットのバッテリー値の推移



c)　1 ロボットの行動の推移

図 3.19: 1 ロボットの情動値、バッテリー値、行動の比較

3.4.4 GA を用いた情動行動学習に関するシミュレーション

シミュレーション②では、情動行動学習のみを行い、適切な行動を自動的に獲得できること確認する。このシミュレーションでは、GA を用いて遷移行列 A と Q の最適化を行った。ただし、Cognition モジュールにおける情動の形成は、3.4.3 節におけるシミュレーションと同様に手動で行った。

情動状態の事前遷移確率行列 Q と行動選択確率行列 A の各パラメータ（計 32 個）を GA によって最適化した。本研究では基本的な GA 手法である Simple GA を用いた。Simple GA では個体を表す染色体は 0 か 1 のビット配列によって構成される。よって本論文では、ある一状態からの遷移を表す 4 つの各遷移確率を 12bit で表現する。染色体の構成図とパラメータ復元の様子を図 3.20 に示す。ある一状態からの各遷移の確率和は 1.0 であるため、各遷移確率は 1.0 を四分割するための三つの小数(全体を分割、左側を分割、右側を分割する割合）で表現できる。よって、一つの小数を 4bit で表現し、パラメータ復元の際は、4bit の 2 進数の数値を 10 進数に変換し、得られた数値を 0 から 1.0 の値域に正規化することによって小数を得られる。遷移元の状態数は 4 であり、Q と A の 2 つの状態遷移行列を表現するために、解の候補である 1 個体は 96bit（12×4×2） で構成される。

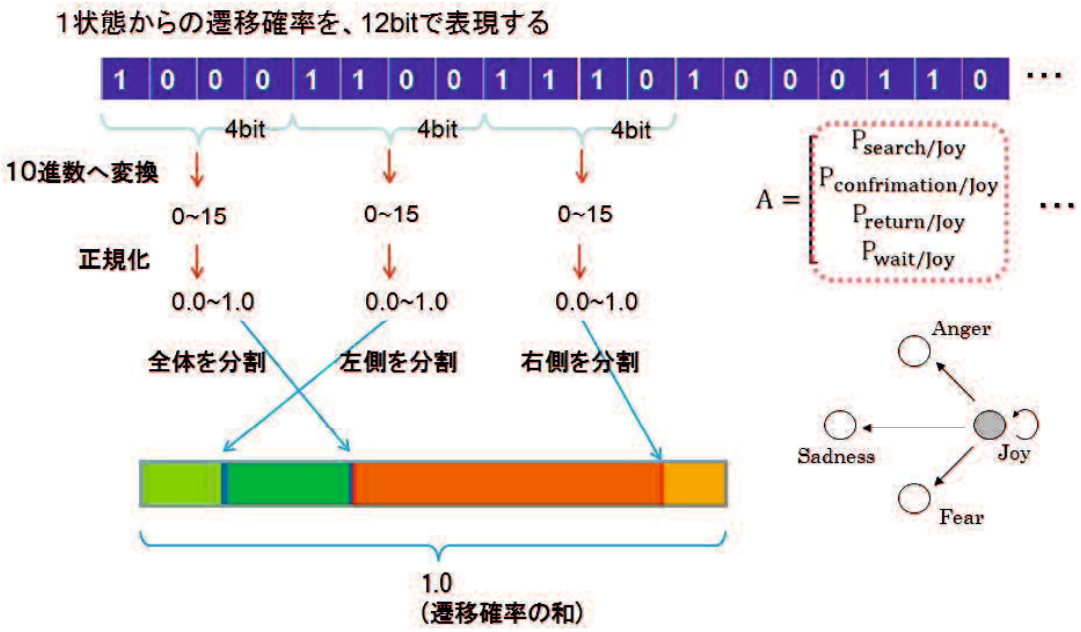
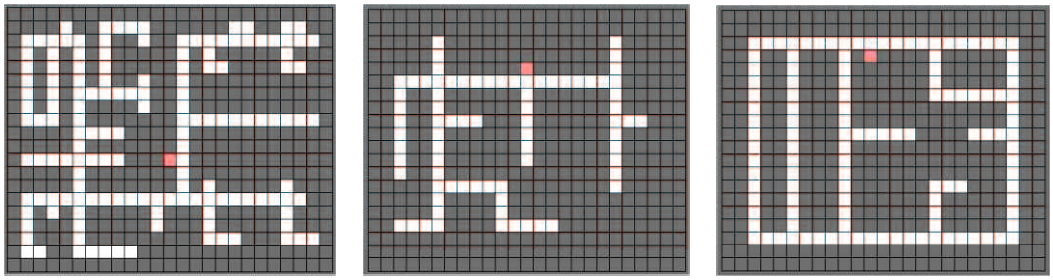


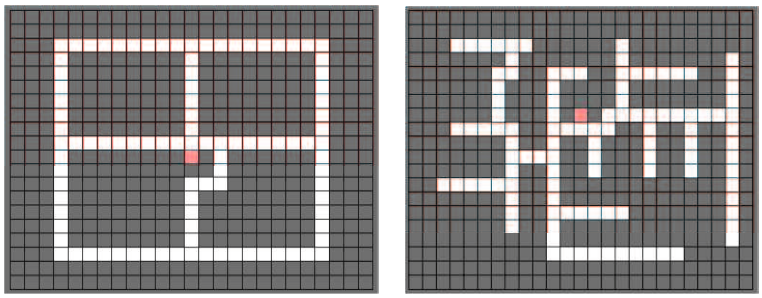
図 3.20: 遺伝子の構成とパラメータへの復元法

適応度を評価するための環境として、構造や広さの異なる 5 つの環境を用意した。使用する環境を図 3.21 に示す。また、遺伝的アルゴリズムの各条件は以下の通りである。

- ✧ 個体数：24
- ✧ 個体の遺伝子の数: 92 (bit)
- ✧ 最終世代：100



(a) Environment 1　(b) Environment 2　(c) Environment 3



(d) Environment 4　(e) Environment 5

図 3.21: 適応度評価で使用する 5 つの環境

適応度 *Fitness* はバッテリー値が 0 ではないロボット Robot 中の *i* 番目のロボットがマッピングしたセル数 $Cells_i$（他ロボットから得たセル情報を含む）とバッテリー消費量 $Consumption_i$、タスクの所要ステップ数 *Step* と最大ステップ数 *MaxStep*（=400）を用いて式(3.18)により決定される。

$$Fitness = \sum_{i \in Robot} \{Cells_i + (150 - Consumption_i) + (MaxStep - Step)\} \quad (3.18)$$

シミュレーションのフローチャートを図 3.22 に示す。1 世代で、すべての個体の適応度の計算と交叉や突然変異を含む世代更新を行う。適応度の計算は、適応度評価用の 5 つの環境全てにつき 2 回ずつ対象タスクを行い、結果より適応度を計算する。

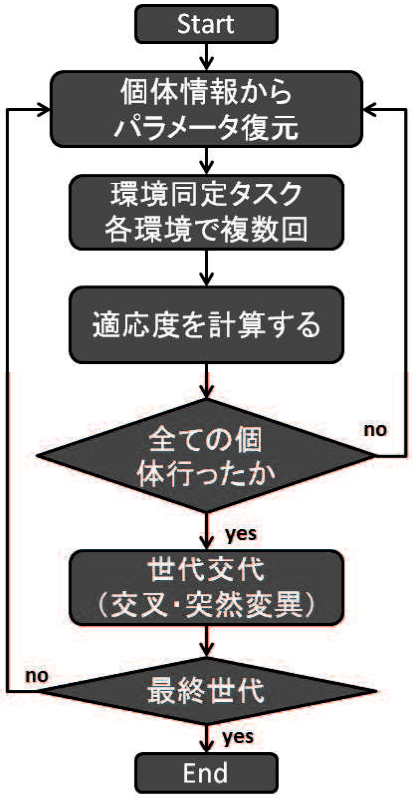


図 3.22: 行動学習のフローチャート

結果における、適応度の世代的変化を図 3.23 に示す。結果より、適応度が高い値へ収束していることがわかる。最終世代における優秀個体の確率行列 Q、A を式(3.18)に示す。GA の効果については、遷移行列 A に関しては全体的な数値の分布は手動時と類似しているが、手動では困難な細かな調整がなされている。例えば$P_{confirm/joy}$と$P_{confirm/anger}$(式(15)における下線部)の値は手動時よりも大きくなり、confirm が選択される確率が高くなっている。次に、手動によるパラメータ設定と、GA によるパラメータ設定の、それぞれ 100 回の試行におけるタスク完了までの所要ステップ数の平均値を比較した。5 つの環境すべてについて比較を行った結果を図 3.24 に示す。図より、いずれの環境においても、GA による最適化を行った方が、手動によるパラメータ設定を用いるよりも少ない所要ステップ数でタスクを終了出来ていることが確認できた。

$$Q=\begin{bmatrix} 0.16 & 0.0 & 0.0 & 0.8 \\ 0.44 & 0.66 & 0.0 & 0.2 \\ 0.0 & 0.11 & 0.53 & 0.8 \\ 0.6 & 0.22 & 0.46 & 0.1 \end{bmatrix}, \quad A=\begin{bmatrix} 0.32 & 0.58 & 0.12 & 0.0 \\ 0.62 & 0.21 & 0.75 & 0.0 \\ 0.0 & 0.2 & 0.06 & 0.66 \\ 0.06 & 0.0 & 0.07 & 0.33 \end{bmatrix} \tag{3.20}$$

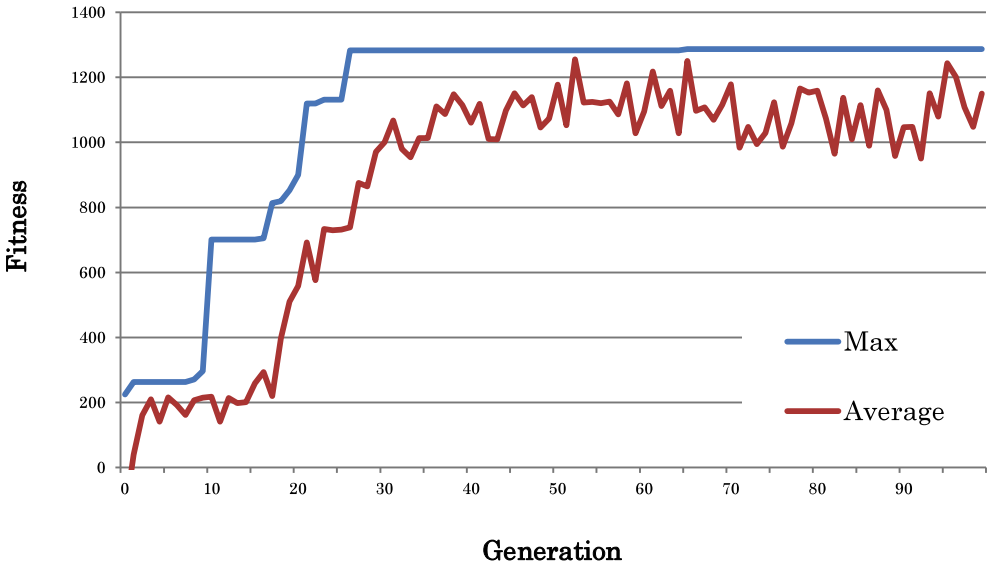


図 3.23: 適応度の世代的変化

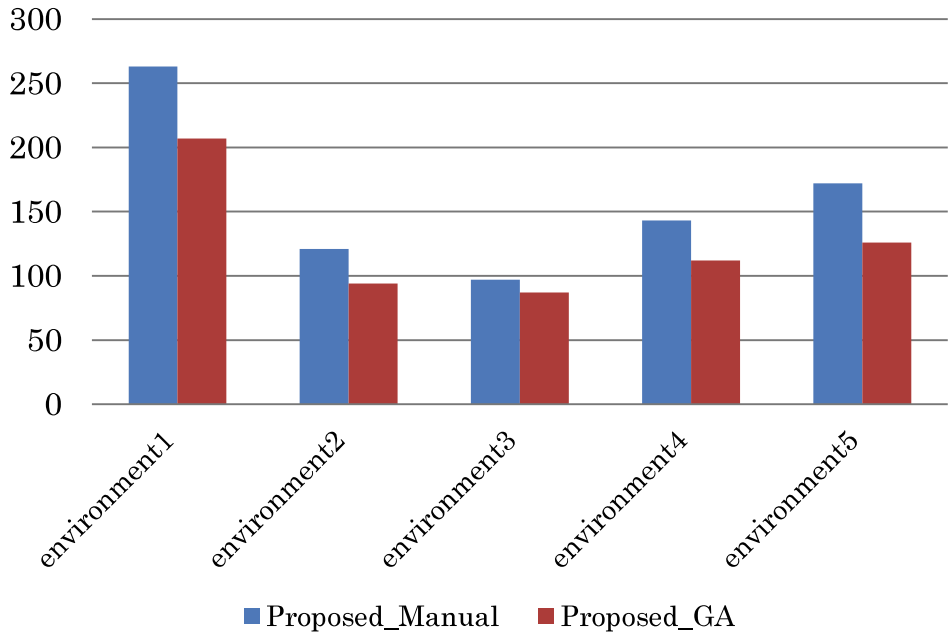


図 3.24: 各環境におけるタスク完了の所要ステップ（100 試行平均）

3.4.5 情動形成学習と情動行動学習に関するシミュレーション

シミュレーション③では、情動形成学習と情動行動学習の両方を行い、感覚刺激に対する情動反応と行動決定が自動的に生成されることを確認する。このシミュレーションでは、提案手法であるニューラルネットワークによる非線形変換の自己組織化能力についてテストし、その後、その提案手法を Cognition モジュールへ導入することにより、システムの性能評価を行う。タスクによって異なる感覚刺激を適切に情動誘発因子へ変換する非線形変換手法についてテストする。

まず、SOM のクラスタリングシミュレーションを行った。これは特定のタスクを想定した場合の感覚刺激の数、値域に従ったランダムな入力値を発生させ、教師無し学習を行う。このときの SOM のパラメータ設定を表 3.1 に示す。本シミュレーションでは構築結果の可視性を重視するため感覚刺激数を 2 とした。また感覚刺激 Z の各値域は次の通りである。

- ✧ z_1: ((*Battery*)-(*Distance from the Base*)) : (0, 100)
- ✧ z_2: (*Number of Fronteir*) : (0, 6)

表 3.1: SOM のパラメータ設定

	SOM
Number of learning	500
Number of nodes	10×10
Learning coefficient η	0.1
$\sigma(0)$ ※ $\sigma(step)=\sigma(0)(1.0-step/\max Step)$	8.0

学習した 2 次元 SOM の z1、z2 それぞれに対応するコードベクトルのマップ変化を図 3.25 に示す。この時マップの X 軸 Y 軸は情動誘発因子に対応しているため、感覚刺激が情動誘発因子の状態に対してクラスタリングされていることがわかる。図では、学習前は乱数により結合荷重が初期化されているため、"バッテリー残量"と"未マッピングエリアの数"のどちらのグラフも不規則な状態である。しかし、学習後は両グラフとも規則性のある状態へ変化している。

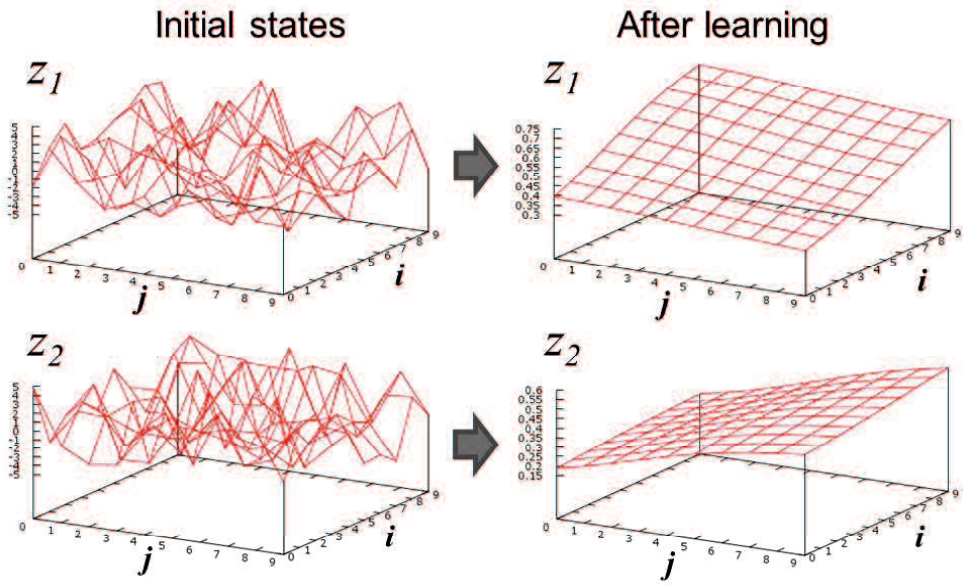


図 3.25: 感覚刺激数が２である場合の、SOM のクラスタリングの様子

次に、学習した SOM のノードを使用して、MLP の誤差逆伝搬法による学習を行った。この時、感覚刺激の数は、"バッテリーの残量"と"未マッピングエリアの数"の２つとした。

この時の SOM および MLP の設定を表 3.2 に示す。

表 3.2: SOM と MLP のパラメータ設定

	SOM	**MLP**
Number of learning	500	10000
Number of nodes	10×10	30 (Hidden layer)
Learning coefficient η	0.1	0.01
$\sigma(0)$ ※ $\sigma(step)=\sigma(0)(1.0-step/\max Step)$	8.0	

学習における、目標出力と近似出力の 2 乗誤差の推移を図 4.24 に示す。結果より、誤差が 0 付近に収束していることがわかる。このことから MLP の近似が正しく行われていることを示す。

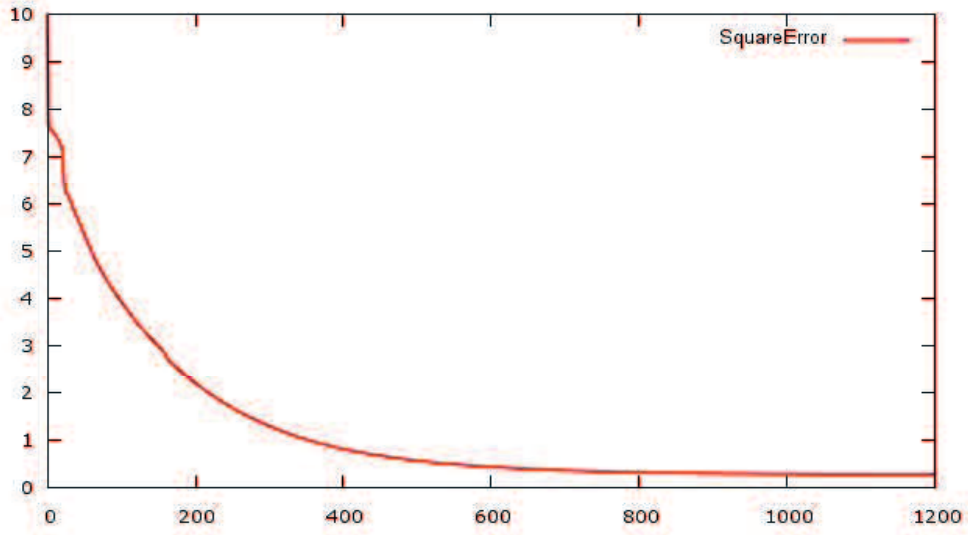


図 3.26: 目標出力と近似出力の 2 乗誤差の推移

学習後の MLP の出力結果について、α、β 別に図 3.26 に示す。このグラフは、MLP へ z_1、z_2 の入力を与えた時の情動誘発因子の出力 α、β の分布を示しており、z_1、z_2 が決まれば α、β が一意に決定されることがわかる。グラフより、生成された非線形変換では、α の値は z_1 に、β の値は z_2 に依存しており、これは、バッテリー値が減少することにより Sadness の情動が誘発され、また、frontier の数が多いほど Anger の情動が誘発されることがわかる。

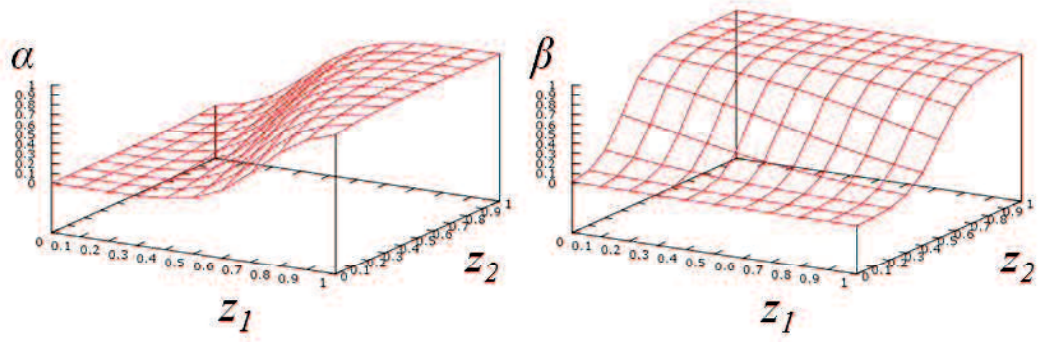


図 3.25: 目標出力と近似出力の 2 乗誤差の推移

最後に、学習が完了した MLP を提案システムの Cognition モジュールに導入し、前シミュレーションと同様に比較を行った。このシミュレーションでは、構築した非線形変換に対して適切な設計パラメータを GA により設定した。

Cognition モジュールに MLP 導入前の Proposed_GA と導入後の Proposed_MLP&GA のそれぞれ 100 回の試行におけるタスク終了所要ステップ数の平均値のグラフを図 3.27 に示す。グラフより、Proposed_MLP&GA は全ての環境について Proposed_GA よりも劣る結果となった。原因として、MLP の学習時に実際の感覚刺激の発生頻度を考慮していないことが考えられる。それにより、状態空間に発生頻度の低い無駄な空間が構築されてしまったため、環境認知能力の低下が起こったと考えられる。しかし非線形変換を自動的に構築したことには大きな有用性があると考えらえる。

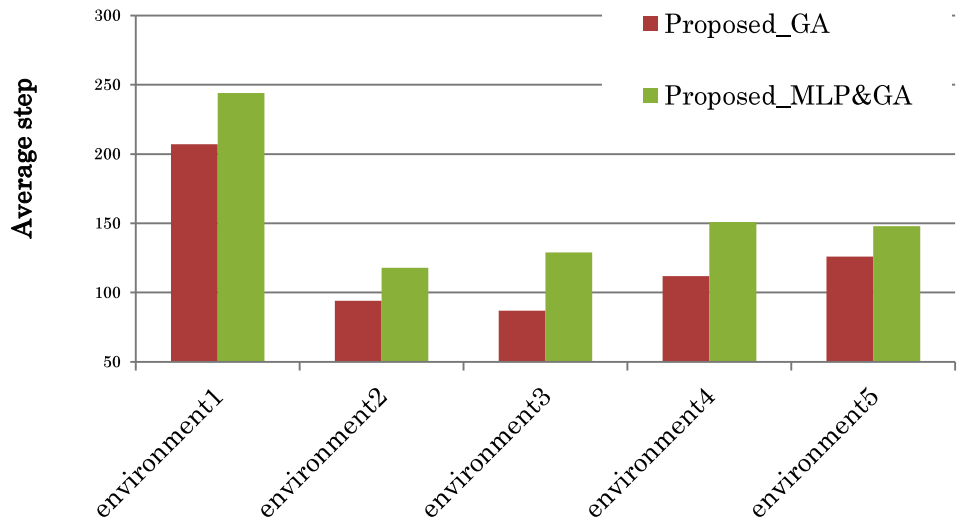


図 3.27: Conventional のアルゴリズム

3.5 考察

本章では、情動に基づく新たなロボットの意思決定システムを提案した。提案システムにおいては情動形成学習と情動行動学習の 2 つの学習により、感覚刺激に対する情動反応のルールと、情動反応に対する情動行動のルールの両方を自動的に生成することを目的とした。提案システムはマルチロボットの環境同定問題を対象とした 3 つのシミュレーションにより評価を行った。シミュレーション①においては、提案システムの性質を調査するために情動形成学習と行動学習は行わず、タスクにおける情動反応や行動決定の振る舞いを観察した。情動ベクトルの値がロボットの状況を反映して変動し、それにより適切な行動が選択された。また、グループ内で各ロボットに適切な役割分担が行われたことが示された。これらの結果より、提案した行動決定手法においては、遷移確率である Q と A の各パラメータを適切に設定することにより、適切な行動決定法を表現可能であることが確認された。次に、シミュレーション②では情動行動学習のみを行い、適切な行動を自動的に

獲得できること確認した。GA によるパラメータの最適化によって、手動によるパラメータ設定よりも性能の良い行動決定法が生成可能であることを確認した。今回は探索時間等を評価の基準としたが、タスクの目的によって様々な最適解が存在する。例えば、より早いタスク終了を望む場合や、より消費エネルギーを抑えたい場合など、優先事項がケースによって異なる。それらのタスクに則した適応度の決定方法を設定することにより、提案手法においては、エージェントのタスクに適切な行動決定法を自動的に設計することが可能である。

最後に、シミュレーション③では情動形成学習と情動行動学習の両方を行い、感覚刺激に対する情動反応と行動決定が自動的に生成されることを確認した。結果より感覚刺激からの情動反応への変換は正しく自己組織化され、MLP によって表現されることを確認した。

しかし、学習した MLP を用いた結果では、従来の Cognition モジュールを備えた提案手法よりも劣る性能評価となった。この原因は、感覚刺激の教師無し学習時に、感覚刺激の発生確率を考慮しなかったため、状態空間に発生頻度の低い無駄な空間が構築されたためだと考えられる。そこで、この課題に対して次章で説明するロボットの経験に基づく情動の再形成によって解決を目指す。

第4章　　ロボットの経験に基づく情動の再形成

本章では、ロボットの経験に基づく情動の再構成に関する研究成果について述べる。まず、4.1 節では情動の再構成の意味と狙いについて述べ、その後、4.2 節では情動の再構成を実現するための提案手法の改良について述べる。4.3 節では移動ロボットのシミュレーションによる提案手法の評価を行う。そして最後に 4.4 節で考察を述べる。

4.1 情動の再形成の意味と狙い

我々の以前の研究では情動の形成を感覚刺激の入力予測値を用いたオフラインでの事前学習によってのみ行った。これにより、実際のタスクにおいてはロボットがあまり経験することがなく参照頻度が低い感覚刺激に関しても情動反応に対応付けられる。限られた情動状態空間を用いて情動行動のルールを記述する提案システムにおいては、このような参照ルールが低い感覚刺激に対する情動反応の対応付けは無駄である。そこで本研究ではロボットが実際にタスク中に経験した感覚刺激のデータを用いた情動の再形成を提案する（図 4.1）。情動の再形成により、適切な感覚刺激と情動誘発因子の対応付けが行われることで、不必要な行動決定法が排除され、重要な行動決定法に充てられる表現量が増加すると考えられる。

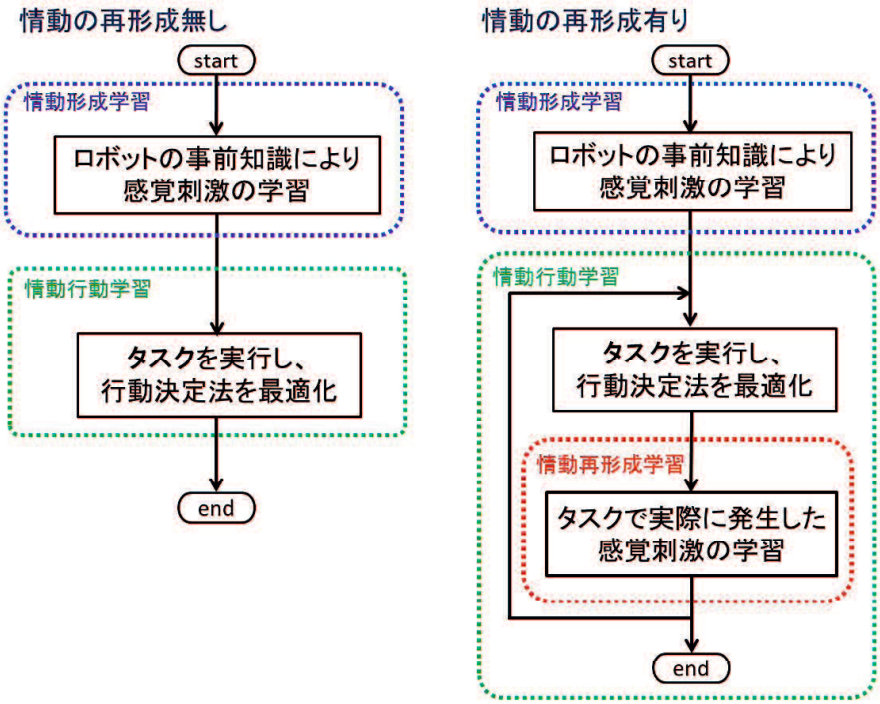


図 4.1: シミュレーションのフローチャート

4.2 情動再形成導入のためのシステムの改良

以前の研究においては、感覚から情動誘発因子への変換を行う Cognition モジュールに MLP を用いており、MLP の学習における教師データの生成に SOM を用いた。MLP を使用することで SOM の各ノードに離散的に保管されたコードベクトルを、連続的に扱うことができた。しかし、MLP は近似能力に優れる一方、学習に多大な反復計算を要し、ロボットが試行中にオンラインで情動状態空間を再構成する場合には不向きである。そこで、MLP を使用せず、SOM のみを用いて情動誘発因子を決定する方法へ Cognition モジュールを改良する。

本研究における Cognition モジュールの構造を図 4.2 に示す。SOM の入力層の各ノードは感覚刺激に対応し、ノード数は感覚刺激入力数 N である．また、従来の Cognition モジュール同様に SOM の競合層におけるノードの位置を情動誘発因子の値に対応させる。つまり Cognition モジュールはロボットが得られる感覚刺激を入力信号として競合層へクラスタリングすることにより、感覚刺激と情動誘発因子との対応付けを構成する。また、感覚刺激を入力信号として決定された勝者ノードの位置情報を 0 から 1 の範囲に正規化することで、感覚刺激を情動誘発因子へ変換することが可能である。

本研究における改良により、Cognition モジュールが出力する情動誘発因子の値が離散的になり、複雑な感覚刺激を扱う場合は性能が低下する恐れがある。しかし出力値の解像度は SOM のノード数に依存するため、SOM のサイズを調整することで性能低下を回避可能であると著者らは考える。

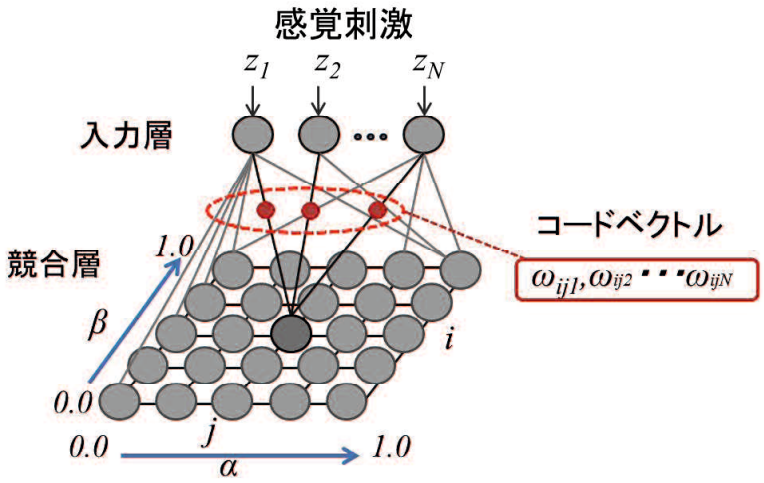


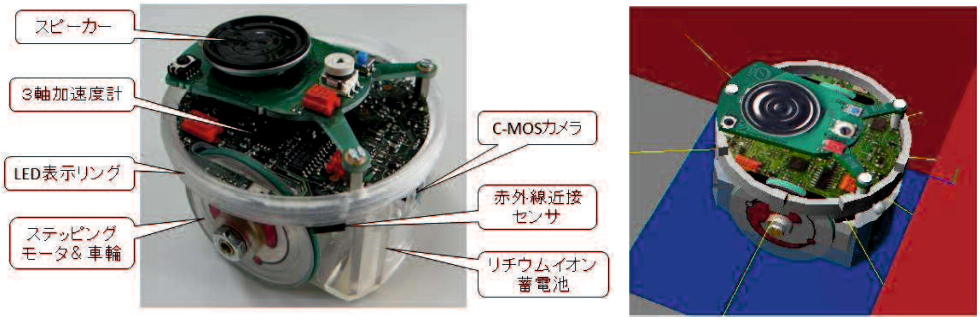
図 4.2: 改良された cognition モジュール

4.3 計算機シミュレーション

提案システムの有用性の検証のために、移動ロボットが通過した床面積の最大化に関するシミュレーションを行った。家庭用掃除ロボットのような環境のマッピングには不十分なセンサーしか搭載していないロボットにおいては、装備するセンサーなどの限られた情報を基に障害物を回避し、環境内のより多くの領域を通過する行動決定アルゴリズムが要求される。そこで本研究では、距離センサーを備えた二輪移動ロボットが一定時間内により多くの領域を通過するための行動決定法が、提案手法を用いて自動構築される様子をシミュレーションにより示す。このタスクにおいては、提案システムにより獲得された行動選択によってロボットがタスクにおいて経験する感覚刺激は変化する。例えば、壁伝いに進む行動決定法を獲得したロボットは、右もしくは左のセンサーに壁を検知した状態に対応する感覚刺激の発生頻度が高い。また、壁に近づくと即座に方向転換を行う行動決定法を獲得したロボットは、多方向を壁に囲まれた状態に対応する感覚刺激の発生頻度が低い。つまり、行動決定法の学習以前には、設計者は Cognition モジュールの学習のために感覚刺激を適切に予測することが困難である。そこで、ロボットは獲得した行動決定を用いたタスクの試行の中で、自身の感覚刺激の経験に基づいて感覚刺激と情動反応の対応づけを適切に修正することが望ましい。

4.3.1 問題設定

提案手法の実装対象である移動ロボットは e-puck とし、性能評価のための実験は、現実に近い環境でシミュレーションが可能な Webots を用いたシミュレーションにより行った. e-puck は研究用として開発・販売されている高機能小型移動ロボットである[70][71]。直径約 7cm、高さ約 5cm の小型ボディにマイクロコントローラ、各種センサー類（C-MOS カメラ、赤外線近接センサなど）を搭載されており、2 つのステッピングモーターを用いることで自律走行ができる。E-puck の外観を図 4.3 に示す。また、e-puck は測定レンジが 4cm の 8 つの赤外線センサーを備えている。2 つの車輪にはステッピングモーターを使用しており、移動距離や回転角の計算が可能である。Webots はスイス連邦工科大学ローザンヌ校(EPFL)で開発されたロボットシミュレータで､現在は Cyberbotic 社が開発･販売を行っている[72]。主に Webots 内に組み込まれている任意のロボットを仮想的に構築し、その動作を確認するのに利用する。シミュレータでは 3 次元 CG で表示し、物理演算を行っているため、実際のロボットに近い動作を実現できる。Webots における e-puck シミュレーションの実行例を図 4.4 に示す。



(a) 実機の外観　　(b) Webots 上での外観

図 4.3: e-puck の外観

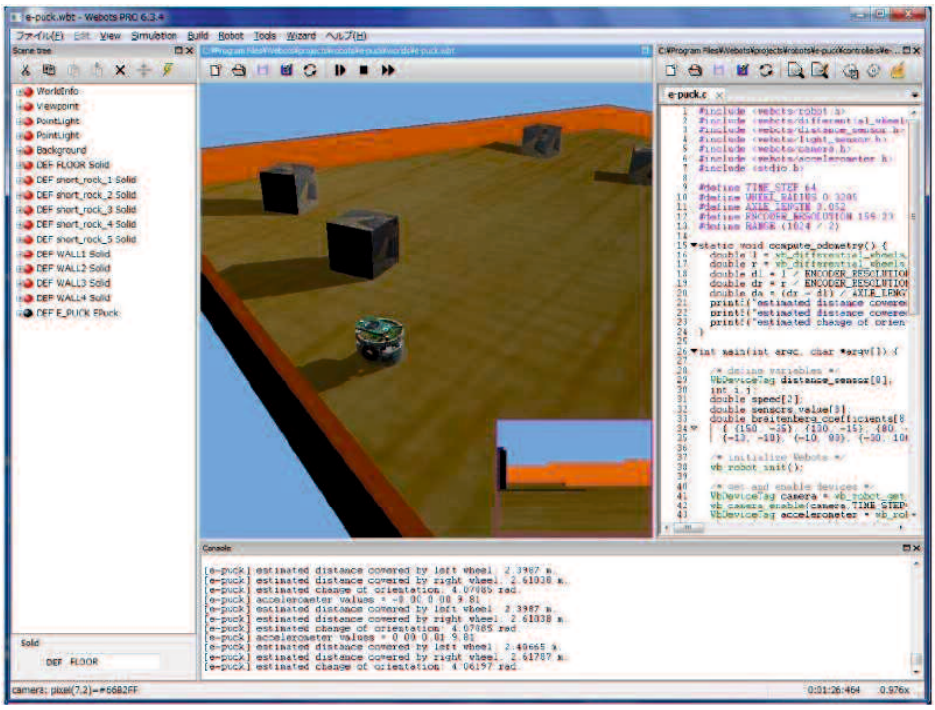


図 4.4 Webots における e-puck のシミュレーションの実行例

Webots シミュレータ上で構造が異なる 4 つの環境を作成し、それらの各環境上で e-puck は移動タスクを行い、通過床面積を最大化する行動決定法を獲得する。図 4.5 は 4 つの環境並べ、その上空から撮影した写真である。 4 つの環境はそれぞれ一辺 100cm の正方形であり、白い領域は壁、黒い領域はロボットが通過できる床である。S はロボットの出発地点を示している。

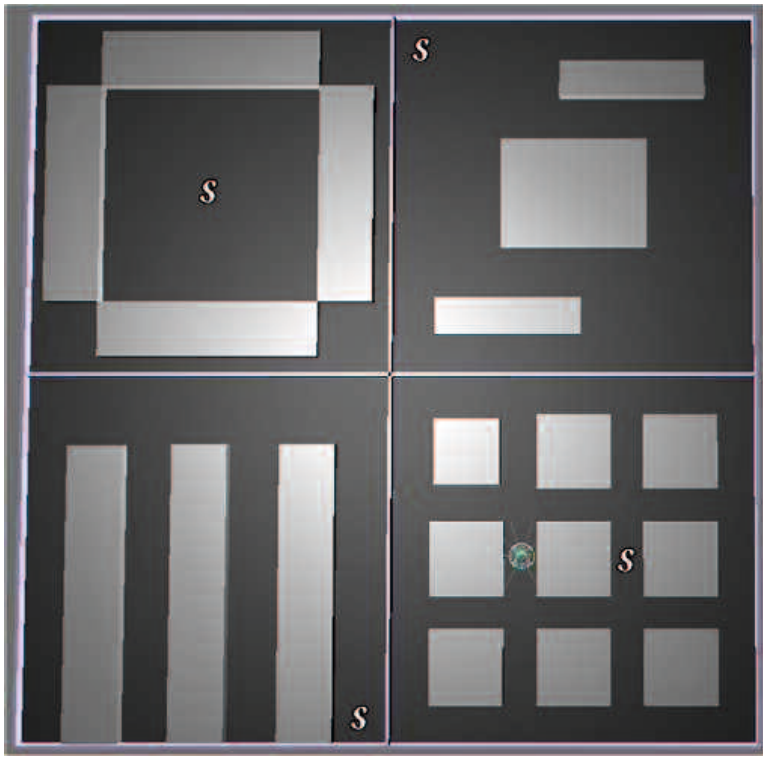


図 5.5 タスクに使用する環境

著者らはロボットの情動行動として直進（drive）、後退（back）、左折(left)、右折（right）４つを定義し、それぞれに対応する左右のモータートルクを表 5.1 のように設定した。

表 4.1: ロボットの行動と対応するモータートルク

行動状態	LeftTorque	RightTorque
drive	100	100
back	-100	-100
left	-50	50
right	50	-50

Behavior Making モジュールでは、式(4.1)式(4.2)のように、各行動に設定された左右のトルクを行動選択確率ベクトル X で重み付け合成することで出力トルクを計算する。

$$\mathrm{MotorTorque}_{left} = \sum_{behavior} LeftTorque_{behavior} \times x_{behavior} \quad (4.1)$$

$$\mathrm{MotorTorque}_{right} = \sum_{behavior} RightTorque_{behavior} \times x_{behavior} \quad (4.2)$$

システムへ入力される感覚刺激数を 6 とし、ロボットの感覚装置との対応を表 4.2 のようにした。また表中のセンサー記号 ps に対応するセンサーの位置を図 4.6 に示す。

表 4.2: システムへの感覚刺激とロボットの感覚装置との関係

感覚刺激	ロボットの感覚装置との関係
z_0	前方 4 つのセンサー距離の平均 (*ps0*、*ps1*、*ps6*、*ps7*)
z_1	右側 2 つのセンサー距離の平均 (*ps1*、*ps2*)
z_2	左側 2 つのセンサー距離の平均 (*ps6*、*ps7*)
z_3	後方 2 つのセンサー距離の平均 (*ps3*、*ps4*)
z_4	単位時間前からのロボットの直線移動距離の累計
z_5	単位時間前からのロボットの回転角の累計

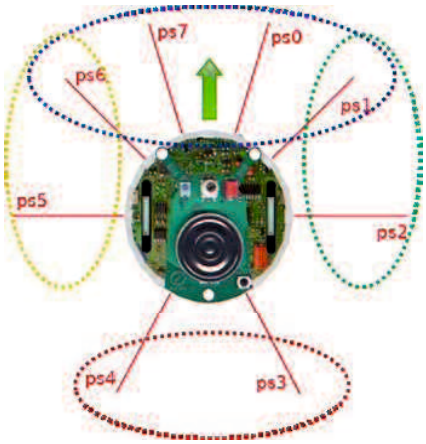


図 4.6: センサーの位置と感覚刺激の対応

4.3.2 感覚刺激の予測値を用いた情動形成に関するシミュレーション

まず SOM による感覚刺激の学習について、事前学習のみを行い、試行による Cognition モジュールの再構成は行わない場合を考える．シミュレーションのフローチャートを図 4.7 に示す。

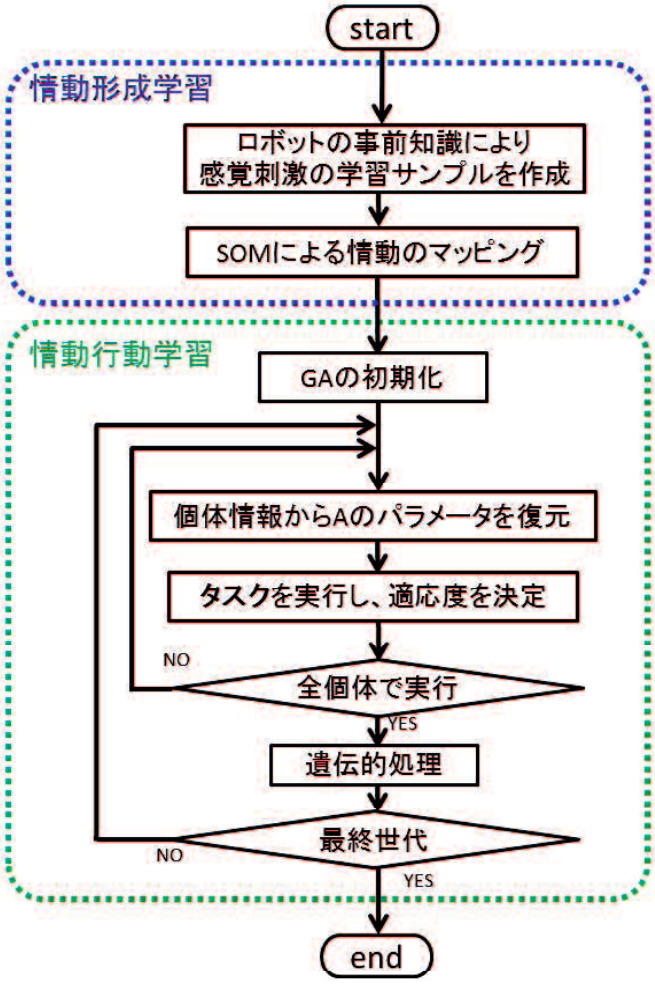


図 4.7: フローチャート

事前学習はセンサーの測定レンジ等の事前知識を基にランダムに作成した学習サンプルを用いる。各感覚刺激の取りえる値域を z_0, z_1, z_2, z_3 は[0:4.0]、z_4, z_5 は[-4.0:4.0]と設定し、ランダムに生成された学習サンプルを用いて 1000 回学習を行った。この時、近傍関数パラメータの初期値 $\sigma(0)$=24.0、学習係数 η=0.1 とした。学習後の SOM の競合層のノードが保管する結合荷重の様子を図 5.8 に示す。6 つの 2 次元マップは SOM の競合層に対応し、感覚刺激別に競合層の各ノードが持つ結合荷重の値を色により表している。すべての感覚刺激の組み合わせが表現されるように色が分布されていることがわかる。これらの 2 次元マップの横軸は情動誘発因子 α に、縦軸は β に対応しているため、感覚刺激の組み合わせを情動状態空間のマップ上に表現することで可読性を高めることが出来る。生成した情動マップを図 4.9 に示す。マップ上の大きな 9 つの円の中には、円が位置する情動を喚起する感覚刺激の組み合わせが表されている。感覚刺激の組み合わせは、障害物との近さを円の大きさ

で、直進距離と回転角を矢印の大きさで表現している。例えば、Joy はロボットの周辺に障害物が少なく直進移動距離が大きい場合に喚起されることが図より読み取れる。

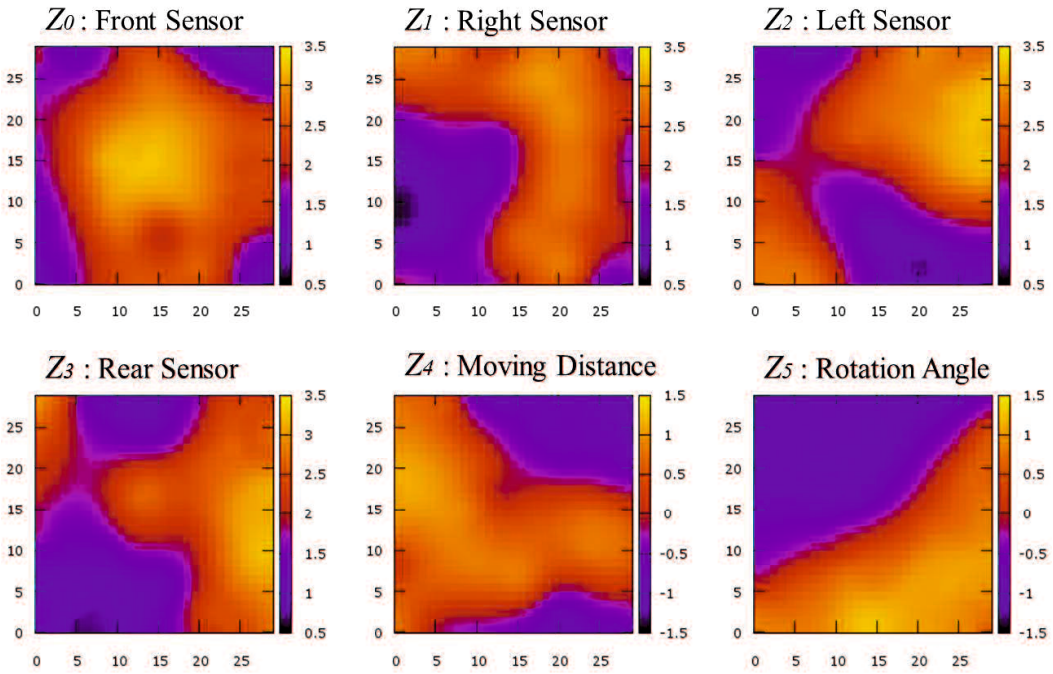


図 4.8: SOM の各感覚刺激に関する値の分布

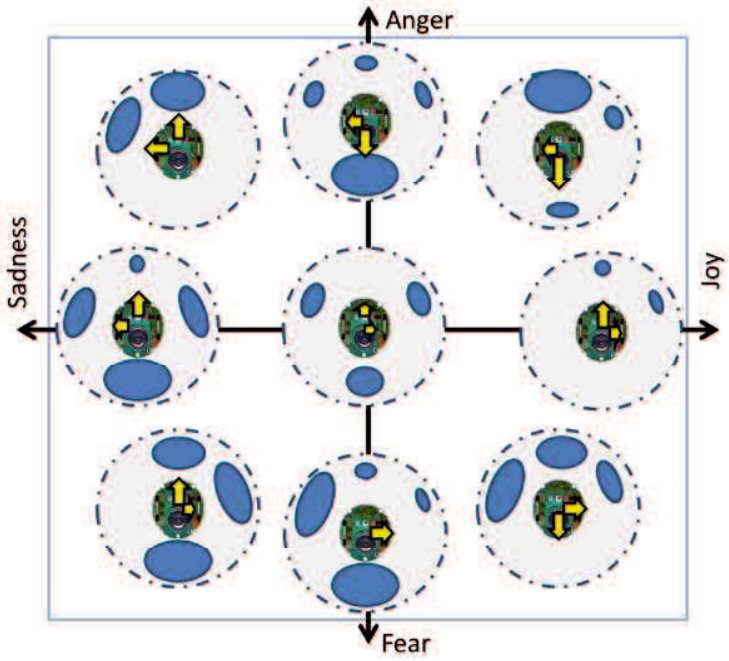


図 5.9 情動マップ

次に事前学習後の SOM を用いて、GA によるパラメータ A の最適化によりロボットの行動決定法を構築した。GA では、各個体において一定時間タスクを実行し、タスク中に通過した床面積を個体の適応度とする。遺伝的進化処理を設定世代数行うとシミュレーション終了とする。まず、シミュレーション環境は図 4.10 に示す障害物が無い狭い環境のみを用いるとし、各個体のタスク開始時にロボットは所定の位置（中央）に設置される。世代数は 25、個体数は 12、1 個体におけるタスクの制限時間は 300s とした。

図 4.10 タスクに使用する環境

GA 実行時における各世代の最大適応度の推移を図 4.11 に示す. グラフより適応度が世代数の増加に従い増加していることがわかる。

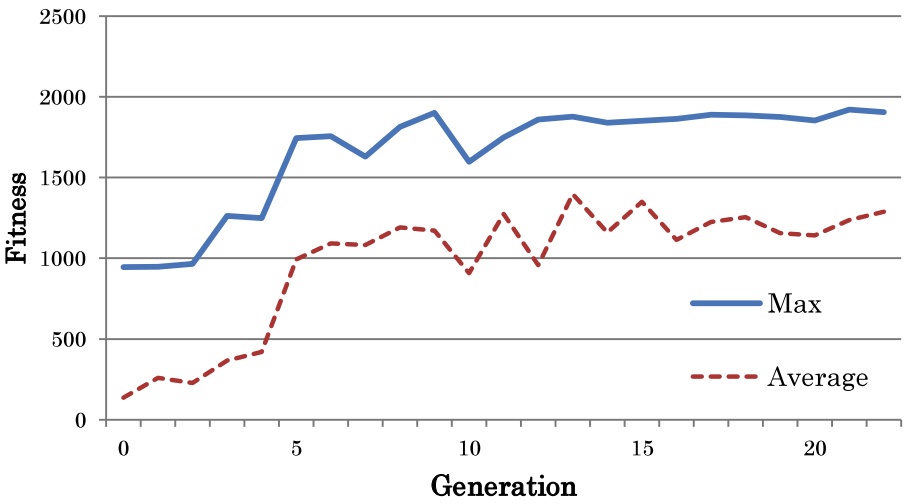


図 4.11: 情動マップ

次に、最終世代における優秀個体のタスクの様子を観察した．環境のロボットの軌跡を白色で示したマップ図を図 4.12 に示す。また、図中には経路の一部を矢印で示している。ロボットは開始直後に少し右に曲り直進した。その後、周辺に壁が無い空間では直進、左前方に壁を検知すると急に右に曲がる、の二つの動作を繰り返した。

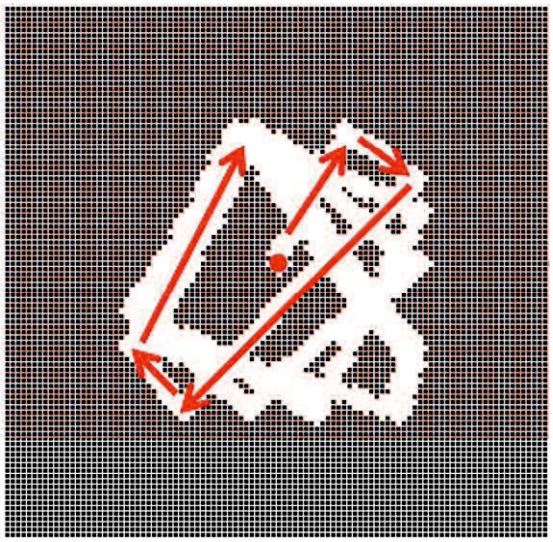


図 4.12 ロボットの移動の軌跡

また、個体のパラメータである確率行列 A を式(4.3)に示す。

$$A=\begin{bmatrix} P_{drive/\,joy} & P_{drive/\,anger} & P_{drive/\,sad} & P_{drive/\,fear} \\ P_{left/\,joy} & P_{left/\,anger} & P_{left/\,sad} & P_{left/\,fear} \\ P_{right/\,joy} & P_{right/\,anger} & P_{right/\,sad} & P_{right/\,fear} \\ P_{bacl/\,joy} & P_{bacl/\,anger} & P_{bacl/\,sad} & P_{back/\,fear} \end{bmatrix}=\begin{bmatrix} 0.80 & 0.13 & 0.00 & 0.40 \\ 0.20 & 0.07 & 0.00 & 0.26 \\ 0.00 & 0.37 & 0.73 & 0.11 \\ 0.00 & 0.43 & 0.26 & 0.22 \end{bmatrix} \quad (4.3)$$

式(5.3)の A では、まず drive の選択確率に関しては$P_{drive/joy}$と$P_{drive/fear}$が大きくなっており、図 5.9 の情動マップを参照すると、Joy は周辺に壁が無い状態（図中の中段右）、Fear は後方と左に壁がある状態（下段中央）に対応している．また、right の選択確率に関して$P_{right/sad}$と$P_{right/anger}$が大きく、Sad と Anger が高い状態は左側と前方に障害物が近い状態（上段左）に対応している．これらによって前述したタスク中のロボットの行動が以下のように説明できる．ロボットは周辺に壁がない状態の時に Joy の情動が誘発され、直進行動の情動行動をとる。また、左側と前方に障害が近づくと Sad と Anger の情動が誘発され、右に曲がる情動行動をとる。

4.3.3 感覚刺激の実測値を用いた情動の再形成に関するシミュレーション

以前の提案システムでは感覚刺激の事前学習により感覚刺激と情動反応の対応付けを形成した。しかし実際のタスク中に発生しない感覚刺激に関しても対応付けが行われる。例えば、ロボットが四方を壁に囲まれている状態はタスク中には発生しない。よって SOM 上に参照されにくい無駄なノード領域が多数発生すると考えられる。そこで実際のタスク中に SOM の追加学習を行うことにより、タスク中に発生しうる感覚刺激の発生頻度を考慮した、より効率的な対応付けが行われはずである。

本シミュレーションでは事前学習後の GA による情動行動の学習部において、序盤の世代のタスク試行中に SOM のオンライン追加学習を導入する。つまり終了世代を 50 世代とし、そのうち最初の 10 世代において、試行中に発生した感覚刺激を用いて情動の再形成を行う。シミュレーションのフローチャートを図 4.13 に示す。まず、ロボットの事前知識により感覚刺激の学習サンプルを生成し、それを用いて SOM の競合層へ感覚刺激のマッピングを行う。その後、GA を用いた情動行動の学習時の実際にタスクの実行により得られた感覚刺激を用いて SOM の競合層へ追加のマッピングを行う。ただし、情動の再形成により既に獲得された行動選択法が無効となる場合があるため、情動の再形成は初期世代のみで行い、その後は情動行動の学習のみを行う。

本実験における事前の情動形成学習は前の実験における事前学習の結果である図 4.8、図 4.9 で表される SOM を使用した。GA の個体数は 24、1 個体の 1 環境におけるタスクの制限時間は 600s とし、図 4.14 に示す 4 つの環境全てでタスクを実行した。各環境での試行における通過床面積の合計を個体の適応度とした。最初の 10 世代において SOM の更新は 1s 毎に行われ、よって SOM の追加学習の回数は 576,000（=600×4×24×10）回である。この時、近傍関数パラメータの初期値 $\sigma(0)=24.0$、学習係数 $\eta=0.0001$ とした。

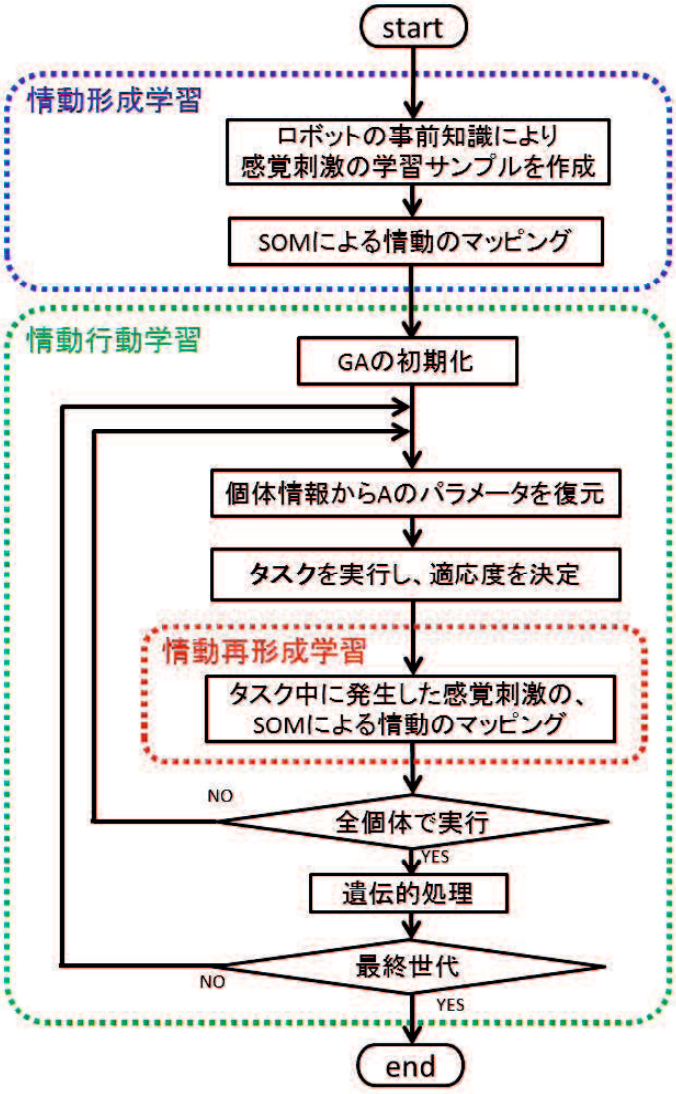


図 4.13: SOM の各感覚刺激に関する値の分布

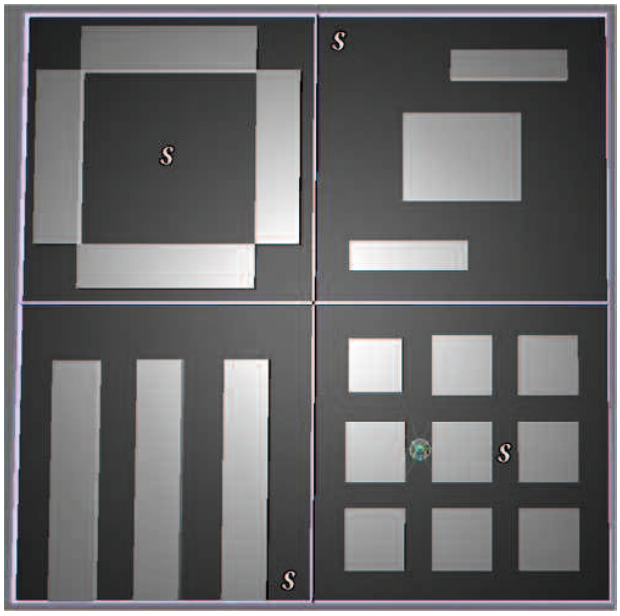
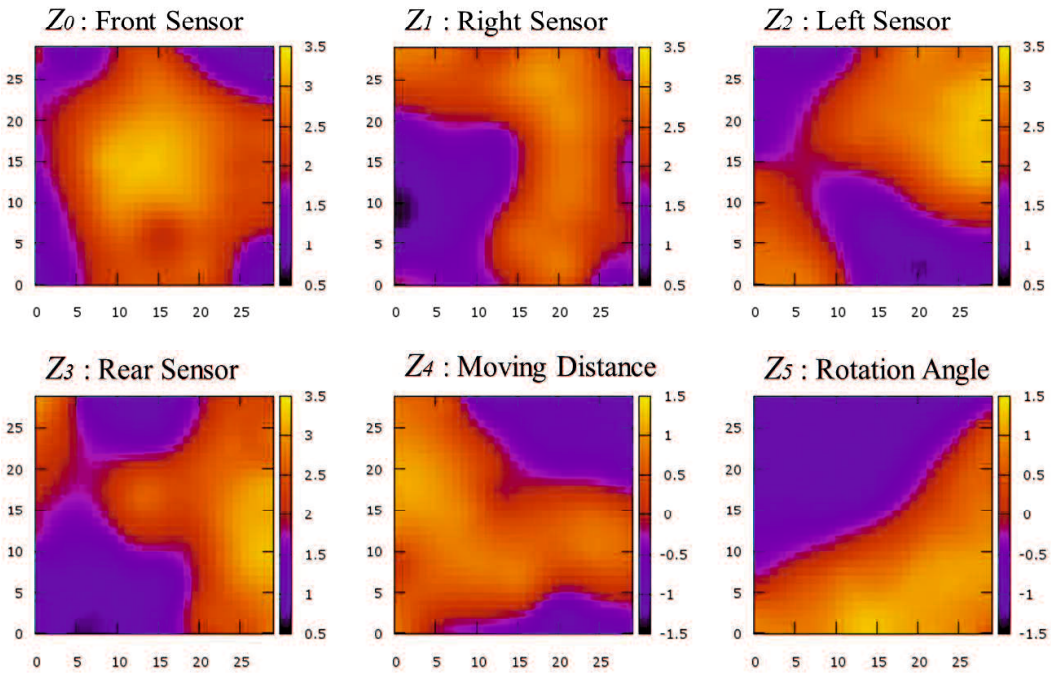
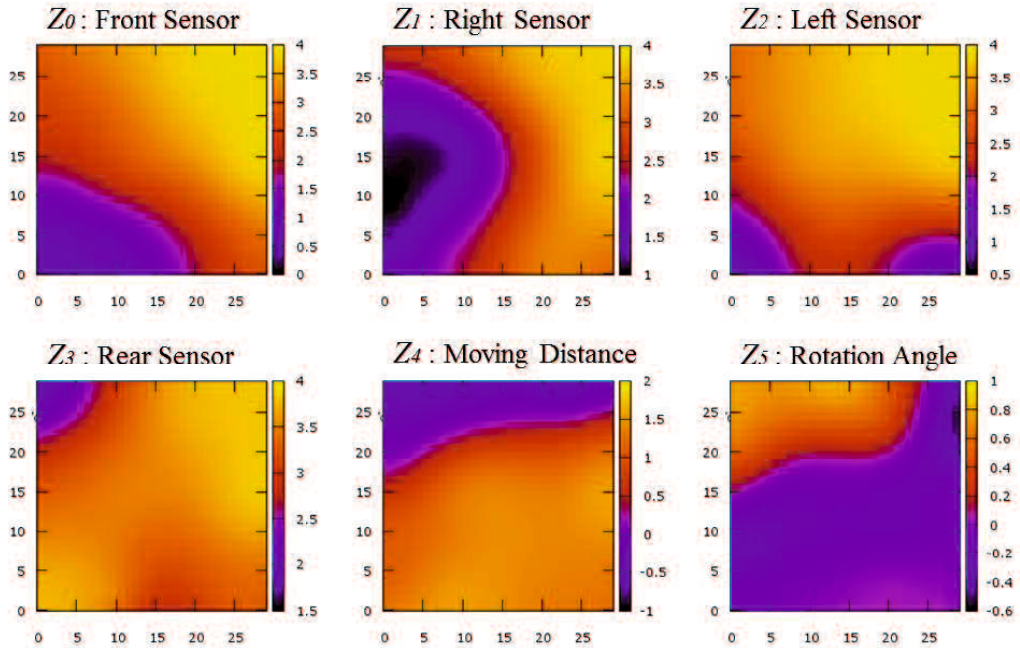


図 4.14: タスクに使用する環境

情動再形成の前と再形成後の SOM の競合層のノードが保管する結合荷重の比較を図 4.15 に示す。結合荷重は各感覚刺激の値を意味しており、図はそれらの分布を表している。再形成後の結合荷重は再形成前と比較して、各感覚刺激に関して値の分布に偏りが生じている。すべての感覚刺激の組み合わせを保有していた追加学習前の SOM に比べて、表現される感覚刺激の組み合わせが減り、よりシンプルになっている。また、図 4.15 の結果を用いて作成した、情動の再形成前と再形成後の情動マップを図 4.16 に示す。再形成後の情動マップにおける Joy や Anger に対応する感覚刺激では周辺に壁が少ない状態に対応しており、タスク中に頻繁に発生される感覚刺激である。また、左側へのカーブや前方への移動が大きい感覚刺激の組み合わせが多く見られる。これらの観察結果より、感覚刺激のタスク中における発生頻度などが SOM に反映されていることがわかる。

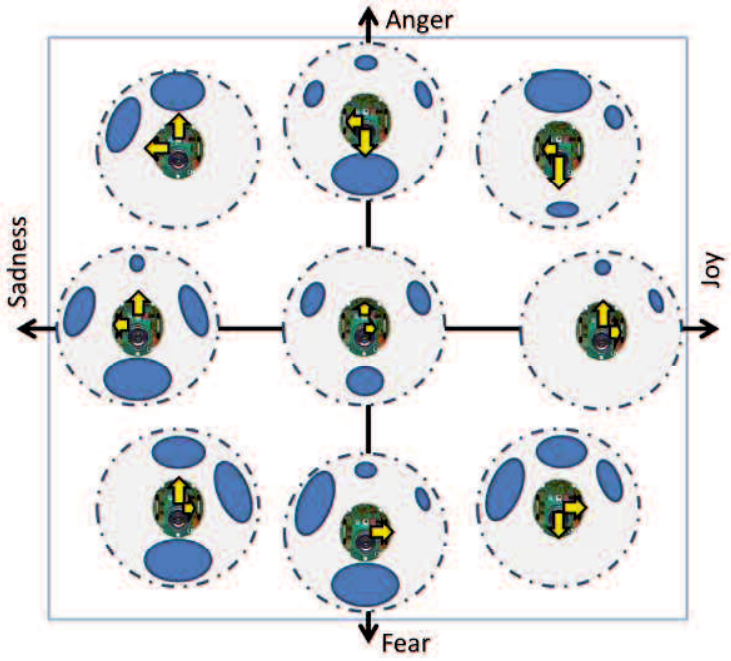


(a) 再形成前

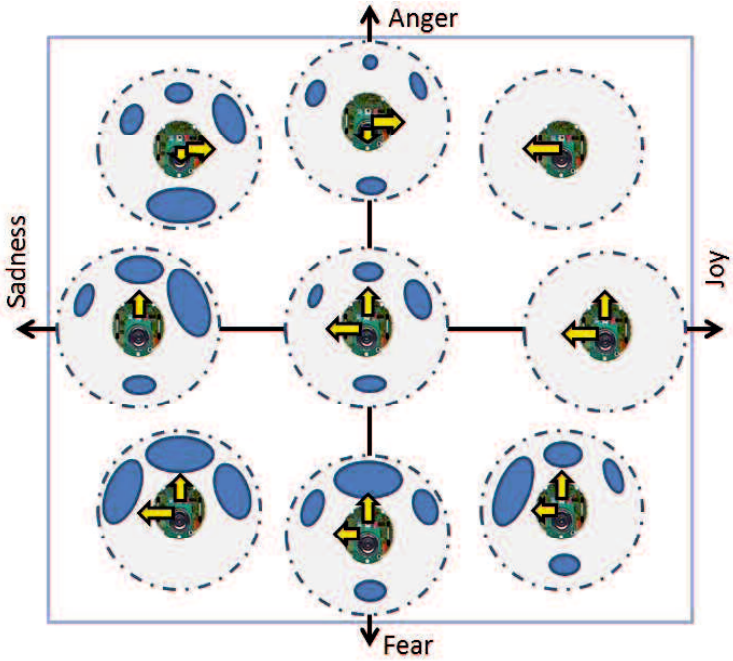


(b) 再形成後

図 4.15: 情動の再形成前と再形成後の SOM の各感覚刺激の分布の比較



(a) 再形成前



(b) 再形成後

図 4.16: SOM の各感覚刺激に関する値の分布

次に環境のロボットの軌跡を白色で表した図を図 4.17 に示す。ロボットはタスク中、壁が無い空間ではやや左に曲がりながら直進し、正面に壁があるときに左へ大きく曲がり、左に壁がある時は右へ大きく曲がった。また、左右を壁に囲まれている状態では直進した。

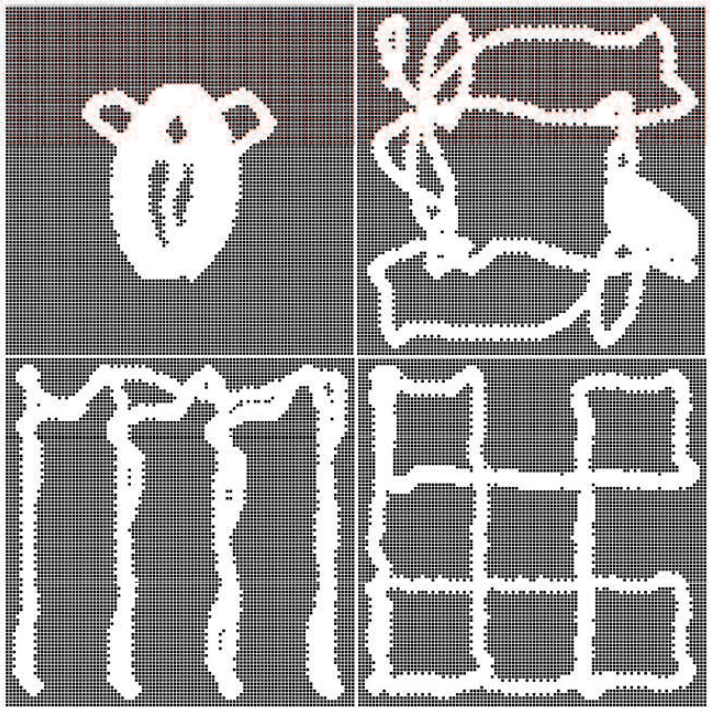


図 4.17: SOM の各感覚刺激に関する値の分布

また、個体のパラメータである確率行列 A を式(4.4)に示す。

$$A=\begin{bmatrix} P_{Dri/joy} & P_{Dri/ang} & P_{Dri/sad} & P_{Dri/fea} \\ P_{Lef/joy} & P_{Lef/ang} & P_{Lef/sad} & P_{Lef/fea} \\ P_{Rig/joy} & P_{Rig/ang} & P_{Rig/sad} & P_{Rig/fea} \\ P_{Bac/joy} & P_{Bac/ang} & P_{Bac/sad} & P_{Bac/fea} \end{bmatrix}=\begin{bmatrix} 0.64 & 0.44 & 0.00 & 0.34 \\ 0.23 & 0.16 & 0.80 & 0.12 \\ 0.12 & 0.21 & 0.00 & 0.42 \\ 0.01 & 0.19 & 0.20 & 0.10 \end{bmatrix} \quad (4.4)$$

シミュレーション①と同様に、A によりロボットの行動決定法が説明できる。例えば、周辺に壁が存在しない状態に対応する Joy からの行動選択確率は Drive が 0.64、Left が 0.23 であり、やや左に曲がりながら直進の動作となる。また、右前方に壁がある状態に対応する Sadness には Left が、左に壁がある状態に対応する Fear には Right の選択確率が高くなっている。

次に再学習を行わない従来の提案システムとの比較を行った。GA における各世代の最大適応度の推移を図 4.18 に示す。ここでグラフ中の垂直線は追加学習が終了する世代数 10 を示している。結果において最終的な適応度は追加学習を行ったシステムの方が高くなった。次に、タスク中に SOM の各ノードが勝利ノードとして参照された回数を調査し、ヒストグラムを作成した。図 5.19 にそれぞれのシステムにおける参照回数のヒストグラムを示す。事前学習のみのシステムでは参照されるノード領域が集中しており、タスク中の情動誘発因子の変化が少ないことがわかる。一方、それに比較すると追加学習を行ったシステムでは参照されたノード領域が分散し、タスク中の情動誘発因子の変化が大きい。つまり、追加学習を行ったシステムではシミュレーション環境内で発生する感覚刺激の変化に情動誘発因子が影響を受けやすい、情動誘発因子と感覚刺激とのより効率的な対応付けが行われたことがわかる。

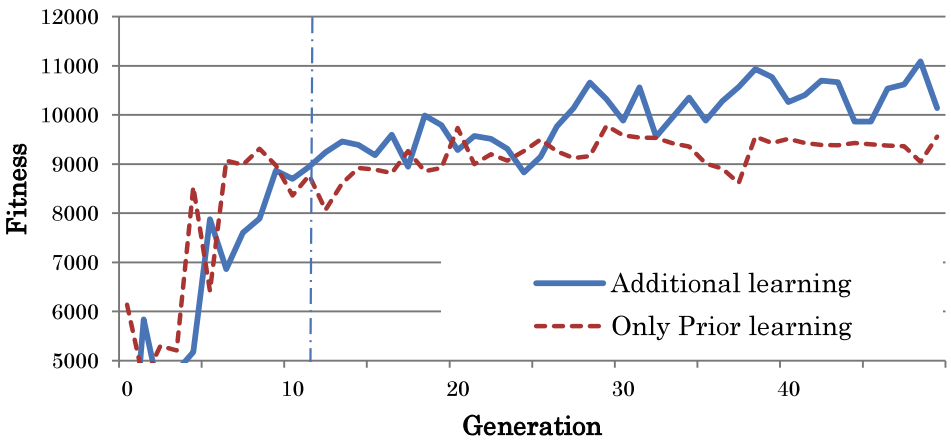


図 4.18: 最大適応度の推移

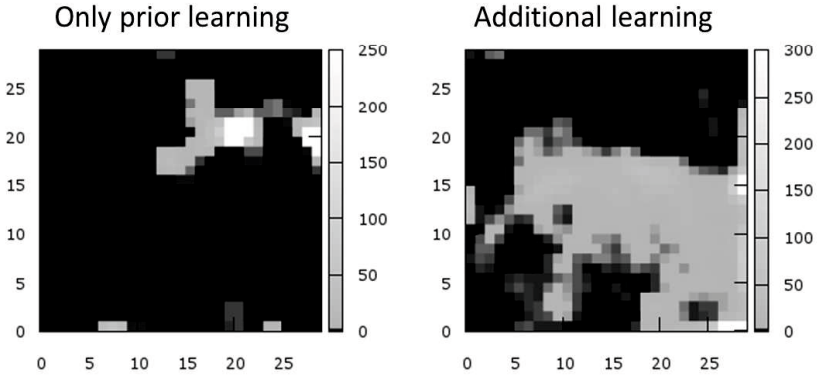


図 4.19: SOM の各感覚刺激に関する値の分布

4.4 考察

従来の提案システムにおいては感覚刺激の教師なし学習時に、感覚刺激の発生確率を考慮しなかったため、発生頻度の低い感覚刺激の状態に対しても情動反応の対応付けが行われた。限られた情動状態空間を用いて情動行動のルールを記述する提案システムにおいては、このような参照ルールが低い感覚刺激に対する情動反応の対応付けは無駄である。また、感覚刺激の発生頻度は獲得した行動選択法によって変化するため、これらを事前に予測することは難しい。そこで本章ではロボットが実際にタスク中に経験した感覚刺激のデータを用いた情動の再形成を提案した。

シミュレーション①の結果では、情動形成学習と情動行動学習の過程により得られた SOM と A の各パラメータを解析することで、情動反応と環境状態との関係性や、その行動決定への影響が意味的に理解可能であることが確認された。また、シミュレーション②の結果では、タスク試行中に実際に得られた感覚刺激により情動を再形成することにより SOM で表現される情動マップ上の参照頻度が低い感覚刺激に対応する領域が消滅し、感覚刺激と情動反応の対応付けが最適化されることより、生成されるロボットの行動決定法の性能を向上させることが示された。

今回は e-puck の実機ではなく、シミュレーションにおいて学習を行った。そのため、パラメータ最適化のためのタスクの繰り返しを現実時間で行うことができた。しかしながら、実際のロボットにおいて進化計算を行うためには多大な時間を必要とする。また、人間が情動のみを用いて行動決定をしているのではないように、提案システムにおいても情動行動学習とは別に基本的な行動決定の学習を行う必要があると考えられる。そこで、提案システムの実ロボットにおける有効性を示すために、次章では提案システムの強化学習への応用を提案する。強化学習による行動学習と提案手法による行動学習の役割を明確に差別化し、実機ロボットの学習においても有効な提案システムの利用方法を提案する。

第5章　　情動進化による強化学習の学習戦略の獲得

これまでの研究ではロボットは情動行動を学習し、情動のみに基づいて行動していたが、本来、人間のように基本的な行動学習は別に行うべきである。また、提案システムにはたくさんの繰り返し計算が必要であり、実機ロボットにおいてこれらを行うことは困難である場合がある。そこで、提案システムの強化学習への応用を提案する。強化学習と情動学習の役割を明確に差別化し、情動行動は強化学習を効率化する学習戦略として利用される。学習戦略はタスクに依存しない効率的な学習のためのルールであり、ロボットの目的タスクの行動学習を加速する効果があるが、学習戦略は目的タスクとは異なるトレーニングタスクを用いて獲得可能である。本章では強化学習による行動学習と提案手法による行動学習の役割を明確に差別化し、実機ロボットの学習においても有効な提案システムの利用方法を提案する。最初に 5.1 節において強化学習の基本知識を述べる。次に、5.2 節では提案システムの強化学習への応用手法について述べる。5.3 節では計算機シミュレーションについて述べ、最後に、5.4 節で考察を述べる

5.1 強化学習

強化学習とは、機械学習の中の一分野であり、試行錯誤を通じて環境に適応する学習制御の枠組みである。観測データに対する解析手順は様々なアルゴリズムで表現されるが、その学習方式は概ね「教師あり学習（supervised learning）」、「教師なし学習（unsupervised learning）」、そしてこの「強化学習（reinforcement learning）」に分類できる。強化学習の基本的な枠組みは図 5.1 のように表される。また、強化学習の流れを以下に説明する。

ステップ１．　学習主体は時刻 t において環境の状態観測 s_t に応じて行動選択を行い、行動 a_t を出力する。

ステップ２．　行動により環境は s_{t+1} へ状態遷移し、その遷移に応じた報酬 r_t を学習主体へ与える。

ステップ３．　学習主体は受け取った報酬 r_t に基づき学習器の状態価値を更新する。

ステップ４．　時刻 t を t+1 に進めてステップ 1 へ戻る。

強化学習は環境に関する事前情報を必要とせず、学習主体が環境との相互作用とその報酬や罰を通して最適な行動方策を学習により獲得していくものである。強化学習は自然界における生体の脳を工学的に模倣した学習システムといえ、生体は報酬を得て罰から逃れるような適切な行動を試行錯誤することによって学習する。未知の環境にも適応できる強

化学習手法は、近年、ロボットの最適な制御手法の学習をはじめ、従来の制御工学で対応できない分野への応用が期待されている。本節では 5.1.1 節で代表的な強化学習手法である Actor Critic 法について、5.1.2 節では強化学習のメタパラメータ制御に関する従来手法について説明する。

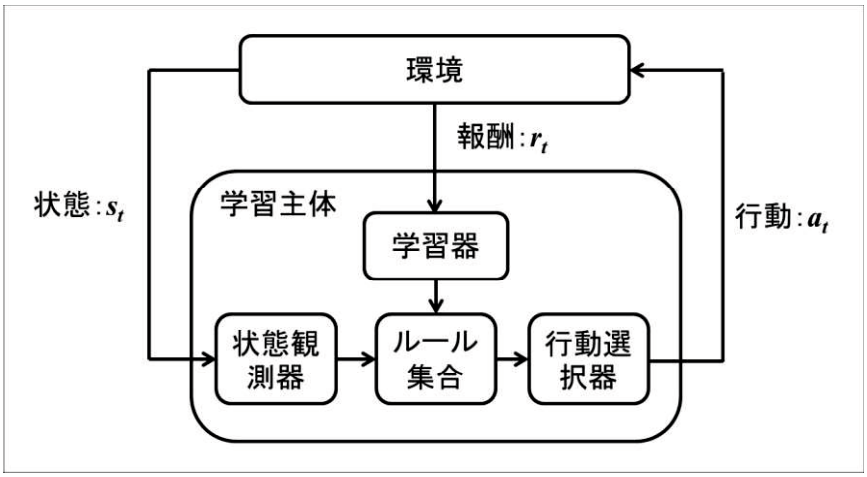


図 5.1: 強化学習の枠組み

5.1.1 ACTOR CRITIC 法

Actor-Critic 法は代表的な強化学習手法であり、状態評価と行動の決定が独立であるという特徴をもつ TD 誤差（temporal difference error）学習手法の一つである。状態を評価する Critic と行動を決定する Actor から構成され、Critic によって求められた TD 誤差から状態評価値および Actor の確率分布を更新する。Actor-Critic の構成を図 5.2 に示す。

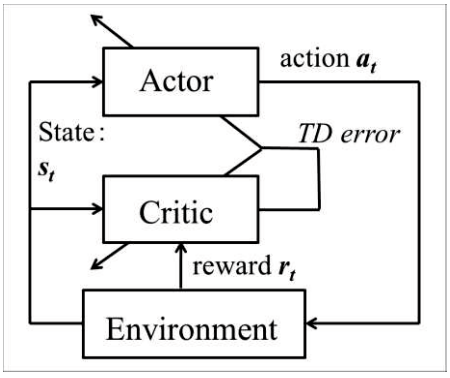


図 5.2: Actor-Critic 手法の全体図

行動選択に最小限の計算量しか必要ないことと、確率的な行動選択を学習できることをメリットとして持つ。Actor-Critic 法の基本的なアルゴリズムを以下に説明する。

TD 誤差は現在の状態の学習主体自身による評価値と、実際に行動することで得た評価値の誤差である。時刻 t における状態を s_t、学習主体による状態の評価値を$V(s_t)$、行動によって得られた報酬をr_{t+1}、またそれにより遷移した状態の評価値を$V(s_{t+1})$として、TD 誤差 *TDerror* は式（5.1）で表される。γは割引率（discount rate）で $0 \leqq \gamma \leqq 1$ の定数であり、現在の行動が将来どれくらい影響を及ぼすかを決めるパラメータである。この TD 誤差が正の時には、見積もっていた評価よりも実際の評価が良かったということであり、負の時には見積もりよりも悪かったことになる。Critic の学習は TD 誤差 $TDerror$ が零になるように、状態価値$V(s_t)$を式(5.2)で更新する。φ は学習係数で $0 \leqq \varphi \leqq 1$ の定数であり、現在と過去を考慮した報酬をどの程度反映させるか決めるパラメータである。また、Actor の学習は状態価値の予測値が高くなるように、行動価値$Q(s_t, a)$を式(5.3)で更新する。また、学習主体の行動の実行の際には、ボルツマン選択により行動選択を行う。ボルツマン選択における行動選択確率は、Actor の行動価値を用いて式(5.4)で決定される。ここで、T は温度定数と呼ばれるパラメータであり、$T \to \infty$の時にランダム選択、$T \to 0$ の時に greedy 選択（最大の行動価値を選ぶ行動選択法）となる。

$$TDerror = r_{t+1} + \gamma \cdot V(s_{t+1}) - V(s_t) \tag{5.1}$$

$$V(s_t) \leftarrow V(s_t) + \alpha \cdot TDerror \tag{5.2}$$

$$Q(s_t, a) \leftarrow Q(s_t, a) + \alpha \cdot TDerror \tag{5.3}$$

$$\pi(s, a) = \frac{exp(Q(s,a)/T)}{\sum_{b \in A} exp(Q(s,b)/T)} \tag{5.4}$$

強化学習には学習式において、事前に設定しなければならないメタパラメータが存在する。Actor-Critic においては学習係数 φ、割引率 γ、温度定数 T がメタパラメータである。メタパラメータの設定は効率的な学習を行う上で非常に重要なファクタである。しかし、その最適設定は学習対象と環境条件に左右されるため、どのような課題、環境に対しても万能に通用するメタパラメータをあらかじめ設定しておくことは一般的には不可能である。そのため、通常、強化学習手法のメタパラメータは経験的に設定され、学習終了まで固定とされている。しかし、外部環境の変化への対応のため、それらを適応的に制御する手法が提案されている。それらを 5.1.2 節で説明する。

5.1.2 環境の変化に順応するためのメタパラメータの制御手法

ある定常的な環境の下で強化学習を用いた場合、試行錯誤を繰り返しその環境に適応した行動を学習していく。しかし、ある程度学習が収束した状態で学習対象である環境やタスクが変化した場合、それまでに獲得した収束行動とは異なった行動が必要となり再学習

が必要になる場合がある。この時、従来の強化学習のアルゴリズムでは既に獲得した収束行動が新たな最適行動の獲得を妨げる恐れがある。この問題に対して、メタパラメータの値を適応的に変化させることにより再学習の効率を改善する手法が提案されている。以下に、神経修飾物質の知見に基づいたメタパラメータ制御法である水野らの手法、溝上らの手法、秋口らの手法について説明する。

水野らが提案している制御法では、現在の報酬が過去に比べてどれだけ減ったかを示すパラメータである *down_rew* に基づいて各パラメータを更新する[55]。*down_rew* は学習中の急激な環境により突然報酬が得られなくなると減少し、次の式(5.7)、式(5.8)に示すアルゴリズムで決定される。

$$\text{if}(down\_rew_{t-1} < down\_rew_t)\ previous\_rew = 0 \tag{5.7}$$

$$down\_rew_{t+1} = down\_rew_t + (current\_rew - previous\_rew) \tag{5.8}$$

ただし、*down_rew* > 0 ならば　*down_rew* =0

各メタパラメータは *down_rew* を用いて式(5.9)、式(5.10)、式(5.11)のように決定される。*height* と *slide* は問題設定に依存するため、その時の環境に合わせて任意に設定する必要がある。

$$\alpha(t) = \alpha_0 \times \left(1 + \frac{height_\alpha}{1+\exp(down\_rew + slide_x)}\right) \tag{5.9}$$

$$\gamma(t) = \gamma_0 \div \left(1 + \frac{height_\gamma}{1+\exp(down\_rew + slide_\gamma)}\right) \tag{5.10}$$

$$T(t) = T_0 \times \left(1 + \frac{height_T}{1+\exp(down\_rew + slide_T)}\right) \tag{5.11}$$

溝上らが提案している制御法は次のように表される。TD 誤差の値に絶対値に依存して変化する変数 *TDerror'* をとり、それに基づいて各パラメータを更新する（*TDerror'*(0)=0）[56][73]。*TDerror'* は式(5.12)で表され、学習初期や再学習の必要が出た時など TD 誤差が大きい時に高くなり、学習が進み TD 誤差が小さくなると *TDerror'* も低くなる。τ は時定数を表す。

$$TDerror`(t) = \left(1 - \frac{1}{\tau}\right) TDerror'^{(t-1)} + \frac{1}{\tau}|TDerror(t)| \tag{5.12}$$

各メタパラメータは$TDerror'$を用いて式(5.13)、式(5.14)、式(5.15)により決定される。ここで、*weight* は$TDerror'$を重みづけするパラメータである。Mizoue らの論文においては、weight の提案はされていないが、TD 誤差の大きさは問題設定の報酬のスケールに大きく依存するため、本研究においては weight を新たに導入する。

$$\alpha(t) = \frac{2}{1+\exp(-TDerror'(t)\times weight_{\alpha})} - 1 \tag{5.13}$$

$$\gamma(t) = \frac{2}{1+\exp(-TDerror'(t)\times weight_{\gamma})} \tag{5.14}$$

$$T(t) = \exp(-TDerror'(t) \times weight_{T}) - 1 \tag{5.15}$$

秋口らは Maslow の欲求階層説に基づき、人間の「生理的欲求」、「安全欲求」、「親和欲求」の３つを独自に定式化し、それぞれロボットの感覚刺激と対応付けた[57]。これらの欲求の度合いを、異なる学習目標をもつ複数の評価値テーブルをもつ目標選択型 Q-Learning における行動選択に用いることで、状況に応じて学習目標を選択的に切り替える適応的な情動行動の学習が可能となった。さらに、銅谷[54]らの仮説に基づき、欲求の度合いを用いて神経修飾物質との関係性からメタパラメータの制御を行っている。これらの研究はロボットに個性を与え、様々な環境条件における適応的な情動行動の獲得が行われており、非常に興味深い。しかしながら、欲求の定式化や情動行動の設計に独自要素が強く、問題設定への依存性が高い。例えば、「餌」の感覚刺激により空腹度（生理的欲求）が減少、「危険領域」の感覚刺激により危険度（安全欲求）が減少、「仲間」の感覚刺激により孤立度（親和欲求）が減少する。そのため、本研究における比較手法としては用いず、参考研究とする。

本研究では、水野らの手法（以下、Mizuno 手法）と溝上らの手法（以下、Mizoue 手法）を比較手法として用いる。

5.2 強化学習の導入のためのシステムの改良

従来の提案システムは、獲得すべき情動行動が目的タスクにおける最適行動を担っていたため、タスクごとに情動形成学習および情動行動学習が必要であった。これらの学習は膨大な繰り返し計算が必要であり、実機ロボットにおいてこれらを行うことは困難となる場合がある。そこで、提案システムの強化学習への応用を提案する。強化学習と情動学習の役割を明確に差別化するために、情動行動は強化学習を効率化する学習戦略（Learning Strategy）の役割を担う。学習戦略はタスクに依存しない効率的な学習のためのルールであ

り、目的タスクとは異なる訓練タスクを用いて獲得可能である。つまり、訓練タスクを用いた提案システムのオフライン学習によって最適な学習戦略を発見し、得られた学習戦略を目的タスクの学習において利用することで目的タスクの学習を効率化する。ここで、タスクとは獲得すべき動作を指す。例えば、人間が特定のスポーツを習得する際に、目的のスポーツとは別のスポーツの習得経験があるほうが目的のスポーツの習得がスムースなケースがある。これは他のスポーツの習得経験がある人は、“ボールの飛距離が短い”という現象に対して“力を強くする”や、“上達しない”という現象に対して“集中する”や“フォームを変えてみる”といった学習戦略をすでに獲得しているためである。このとき既に習得経験があるスポーツが訓練タスクを意味し、それによって得たスポーツ全般の習得に共通する知識は学習戦略、新たに習得を試みるスポーツは目的タスクを意味する。従って、目的タスクが実機ロボットを用いた行動学習であっても、提案システムは訓練タスクとして目的タスクと類似した条件の計算機シミュレーションにより学習戦略を獲得し、それを実機ロボットの学習に用いることで行動学習を効率化が可能である。

次に、強化学習への導入のための提案システムの改良について述べる。改良型の提案システムの全体構造を図 5.3 に示す。システムの全体構造において、強化学習は従来の Robot Controller に置き換えるように導入されている。また、従来の Emotional Behavior モジュールの行動 X を学習戦略とし、モジュールの名前を Learning Strategy とした。

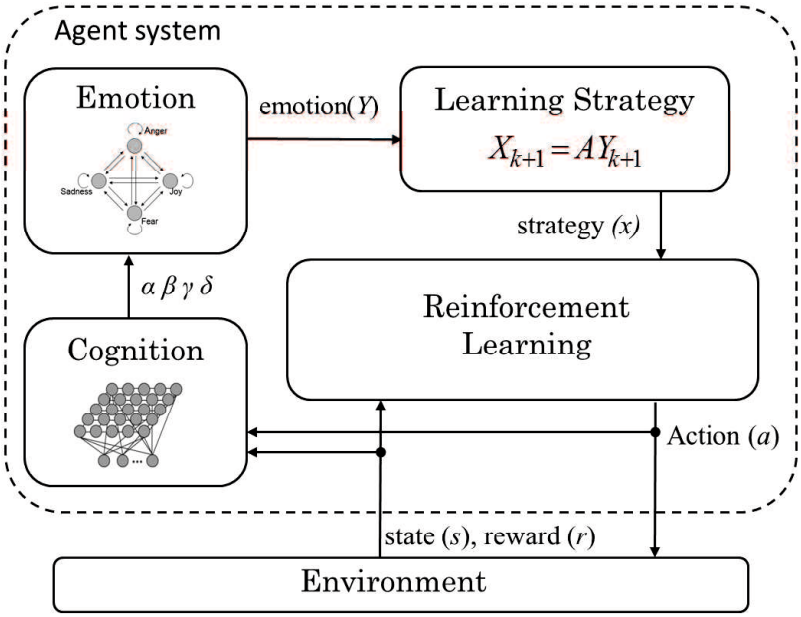


図 5.3: 強化学習を導入した改良型の提案システム

本研究では、学習戦略は具体的に強化学習のメタパラメータの適応的な調節として表現される。これは、人間の「いいかげんに行動する」や「集中する」といった戦略に近いと考えられる。また、異なる考え方では、学習戦略として行動選択の戦略を表現することも考えられる。これは、人間の狭い場所では「動作速度が遅い行動を選択しやすい」といった戦略に近いと考えられるが、本研究ではこちらは採用しない。

提案手法における学習戦略は実際の強化学習タスクを通して行われる。学習戦略の獲得のフローチャートを図 5.4 に示す。

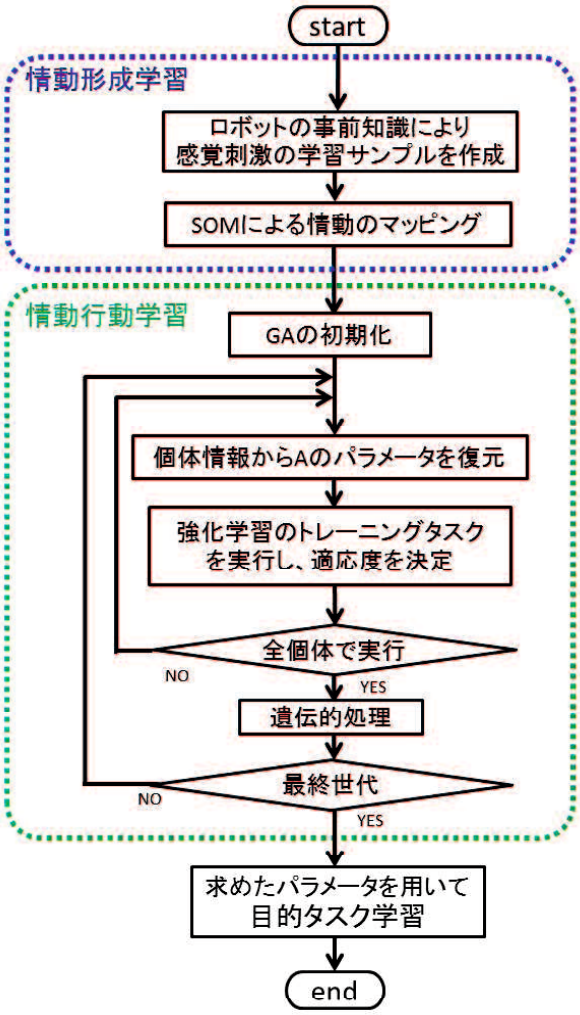


図 5.4: 学習戦略獲得のフローチャート

5.3 計算機シミュレーション

強化学習の迷路探索問題に関するシミュレーションによって提案システムの評価を行った。まず 5.3.1 節では問題設定について説明する。次に 5.3.2 節でシミュレーション結果について述べる。

5.3.1 問題設定

強化学習はロボットの行動学習やスケジューリング問題など様々な分野に応用されているが、様々な強化学習手法の学習性能を定量的に評価するためには実装が容易で明白な性能評価が可能なベンチマーク問題が必要である。そこで、多くの強化学習手法の研究で利用される代表的なシミュレーション問題として迷路探索問題がある。

この問題は格子状のマスにより表現された環境において、所定のスタート地点からゴール地点までのエージェントの最適な移動行動を学習するものである。環境の例を図 5.5 に示す。環境は、エージェントが移動できる通路、もしくは移動できない壁のマスによって構成される、エージェントは 1 ステップで現在位置するマスからその隣接するマスへ離散的に移動する。エージェントの知覚状態（入力状態）は位置とする場合や、エージェント近傍状態とする場合など様々である。一般的にエージェントの移動や壁への衝突に対して負の報酬を与え、ゴールへの到達に対して正の報酬を与えることで、エージェントは試行を繰り返すことによって最適な経路を発見することがきる。エージェントがスタートを出発してゴールまで到達するまでを 1 エピソードとする。エージェントがゴールに到達する、もしくはステップ数が所定の最大値まで達した場合に現エピソードが終了し、エージェントはスタート地点に戻され次のエピソードとなる。また、結果のばらつきを抑えるために、1 試行中の全エピソードが終了するとすべての学習値をリセットし、この試行を数回繰り返して結果の平均値をとる。

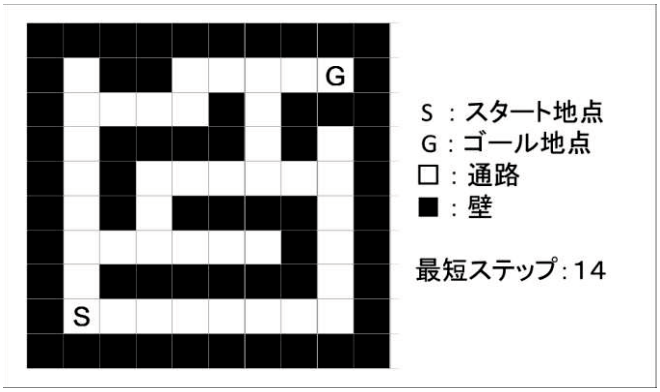
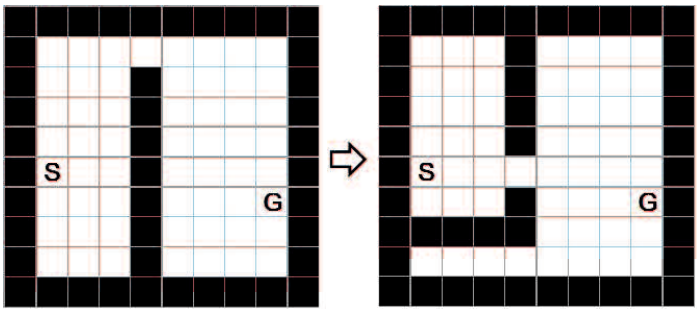


図 5.5: 強化学習を導入した改良型の提案システム

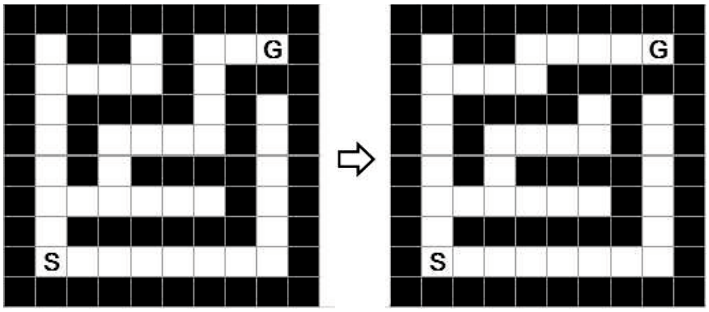
本研究では、環境変化を有する迷路探索問題をタスクとして使用する。この問題ではエージェントの環境学習の途中で探索対象である環境の構造が変化するため、メタパラメータの適応的制御が重要となる。訓練タスクを図 5.6 に示す。このタスクでは、300 エピソード目に環境変化が発生する。初期環境の最適経路のスタートからゴールまでのステップ数は 16 であり、環境変化後の新たな最適経路のステップ数は 8 である。このタスクでは、環境変化前と環境変化後のゴールまでの経路群に共通の状態が存在するため、環境変化前に得られた学習知識を変更後に転移可能である。よって、変化後の環境におけるスタート地点を囲む部屋を出ることが新たな最適経路を発見する鍵となる。ただし、メタパラメータの設定が適切でない場合、変化後にも変化前の学習知識に引っ張られて準最適解に学習収束する可能性がある。



(a) 0~299 episodes (b) 300~499 episodes

図 5.6: 訓練タスク

目的タスクを図 5.7 に示す。このタスクでは、200 エピソード目に環境変化が発生する。初期環境の最適経路のスタートからゴールまでのステップ数は 14 であり、環境変化後の新たな最適経路のステップ数も 14 である。環境の構造は、広い空間で構成された訓練タスクと異なり、細い通路となっている。変更前から変更後に転移可能な知識量が少なく、変更後の環境におけるゴールまでの経路を再学習することが困難である。



(b) 0~199 episodes (b) 200~499 episodes

図 5.7: 目的タスク

これらのタスクにおけるエージェントの認知可能な環境状態 s は、エージェントの座標のみとする。エージェントが受け取る報酬 r は、移動報酬が-0.01、壁への衝突に対する報酬が-0.1、ゴール到達報酬が+10 とした。エージェントの行動 a は上下左右の 4 方向に隣接するマスへの移動とした。

これらのタスクを実行する強化学習手法は Actor-Critic とした。また、提案システムは学習係数 φ と温度定数 T に関する学習戦略を提供し、学習戦略を表 5.1 に示す。これらの学習戦略に従って、各メタパラメータは式(5.16)、式(5.17)のように変動する。

表 5.1: 学習戦略の設定

学習戦略	メタパラメータの変更
strategy1	α を増加させる
strategy2	α を減少させる
strategy3	T を増加させる
strategy4	T を減少させる

$$\alpha \leftarrow \alpha + 0.1 \times x_{stra1} - 0.1 \times x_{stra2} \tag{5.16}$$

$$T \leftarrow T + 10.0 \times x_{stra3} - 10.0 \times x_{stra4} \tag{5.17}$$

比較実験としてメタパラメータの制御の従来手法を用いたシミュレーションを行う。メタパラメータ制御の従来手法として、溝上らの手法（以降、mizoue 手法）、水野ら手法（以降、mizuno 手法）、そしてパラメータを固定した通常の Actor Critic 法（以降 Fixed）を用いる。また、提案手法と同様に、従来手法の各設定パラメータを GA によって最適化した。

この時、個体評価のための適応度は、個体毎に訓練タスクである迷路探索問題を行った結果から計算する。提案システムおよび各従来手法に関する、強化学習および GA の共通設定を表 5.2 に示す。また、各従来手法における GA によるパラメータ設定のフローチャートを図 5.8 に示す。

また、GA は実数値 GA を用い、各個体の評価は式(5.18)で示される適応度関数を用いる。この適応度関数は全エピソードにおいて総ステップ数が少ないほど適応度が高いことを意味している。GA により、提案手法では遷移行列 A の各パラメータを最適化し、各従来手法では表 5.3 で示される設計パラメータを最適化する。

表 5.2: 各従来手法の設定パラメータとその探索領域

強化学習に関する設定		GA に関する設定	
試行回数	200	個体数	200
エピソード数	500	子の生成数	10
最大ステップ数	100	終了世代	3000

$$Fitness = \frac{\sum_{TRIAL}\sum_{EPISODE}(MAX_{STEP}-step)}{(MAX_{TRIAL})\times(MAX\_EPISODE)} \tag{5.18}$$

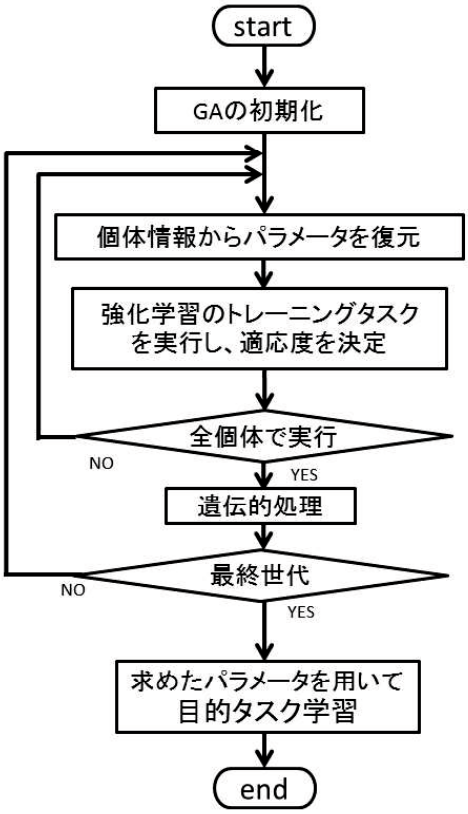


図 5.8: 各従来手法におけるパラメータ最適化のフローチャート

表 5.3: 各従来手法の設定パラメータとその探索領域

Method	Parameter	Range	
		Min	Max
Fixed	φ	0.10	1.00
	γ	0.10	1.00
	T	0.10	1.00
Mizuno's method	$height_{\varphi}$	0.00	1.00
	$slide_{\varphi}$	0.00	4.00
	$height_{\gamma}$	0.00	0.10
	$slide_{\gamma}$	0.00	4.00
	$height_{T}$	0.00	100.00
	$slide_{T}$	0.00	4.00
Mizoue's mehotd	$weight_{\varphi}$	0.01	10.00
	$weight_{\gamma}$	0.01	10.00
	$weight_{T}$	0.01	10.00

5.3.2 メタパラメータ制御法の獲得に関するシミュレーション

本シミュレーションでは、情動の形成はタスクを実行する前にのみ行い、情動の再形成は行わない。まず、感覚刺激の予測値を用いた情動形成学習を行った。SOM のパラメータ設定を表 5.4 に示す。本シミュレーションでは感覚刺激として、メタパラメータ制御に関する従来手法においても用いられる$TDerror'$と Down reward の 2 つの情報を用いた。感覚刺激に対応する入力とその予測される値域を表 5.5 のように設定する。

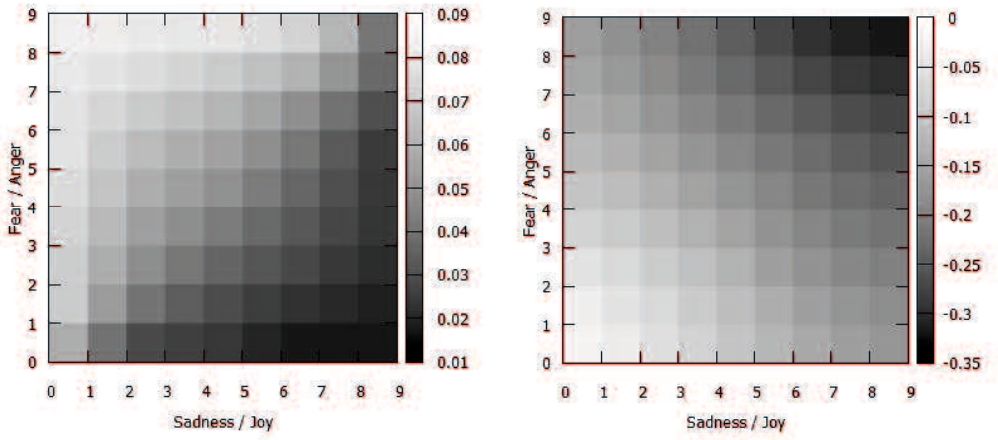
表 5.4: SOM のパラメータ設定

	SOM
Number of learning	500
Number of nodes	10×10
Learning coefficient η	0.1
$\sigma(0)$ ※ $\sigma(step)=\sigma(0)(1.0-step/\max Step)$	8.0

表 5.5: システムへの感覚刺激とロボットの感覚装置との関係

感覚刺激	入力	Min	Max
z_0	$TDerror'$	0.0	0.1
z_1	Down reward	-0.4	0.0

情動形成学習の結果を図 5.9 に示す。例えば、$TDerror'$は Anger と Sadness に対応付けられており、$TDerror'$が上昇すると Anger と Sadness の情動が誘発される。本シミュレーションにおける提案システムでは、この情動形成学習により得られた Cognition モジュールを用いる。



(a) z_1 : $TDerror'$　　(b) z_2 : Down reward

図 5.9: SOM の各感覚刺激の分布

次に、訓練タスクを用いた GA による各手法のパラメータ最適化の結果について述べる。この時、提案手法における情動反応及び学習戦略の実行はエピソード毎に行われ、エピソード中の感覚刺激の平均値によりエピソード終了時に情動反応が起こり、学習戦略によりメタパラメータが変更される。各手法における各世代の個体群の平均適応度の推移を図 5.10 に、個体群の最大適応度の推移を図 5.11 に、個体群のばらつきの推移を図 5.12 に示す。平均適応度を見ると、いずれの手法も世代数が増えると適応度が上昇し収束している。平均適応度においては提案手法の収束値はいずれの従来法に比べても低い。しかし、最大適応度では従来法と同等である。さらに、ばらつきに関しては、従来法に比べて大きい。これは、提案手法は従来法に比べてパラメータの数が多いため、個体群の収束が従来手法に比

べて遅いためだと考えられる。一方、パラメータ数が少ない Fixed と Mizoue は収束が早いと考えられる。

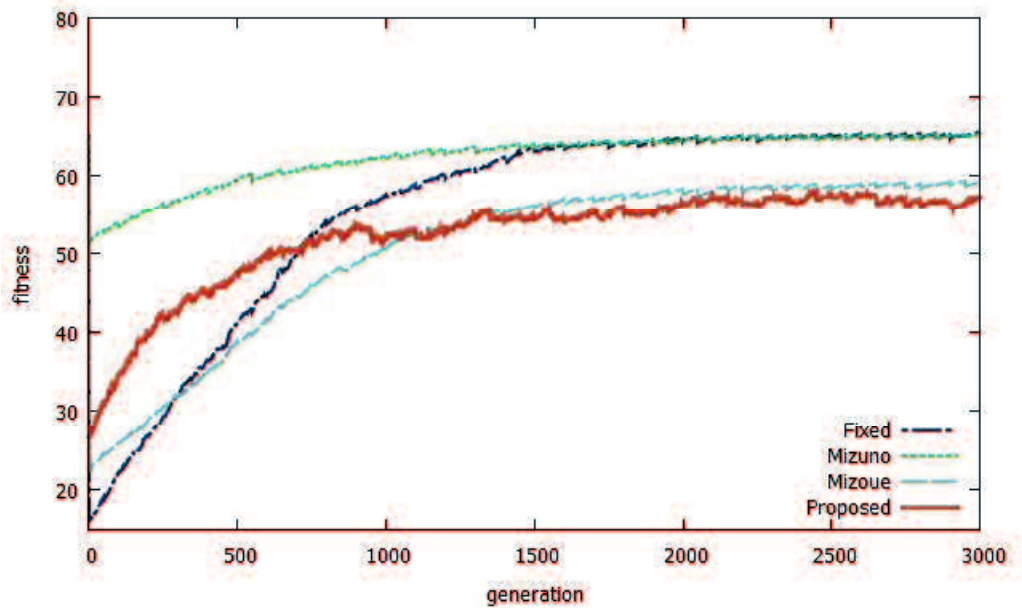


図 5.10: 平均適応度の推移

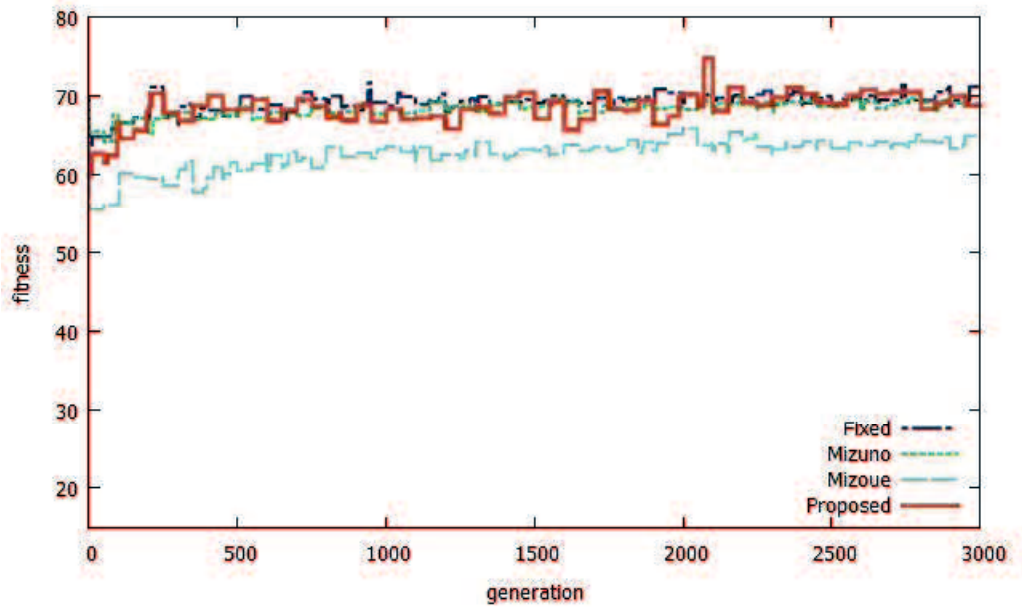


図 5.11: 最大適応度の推移

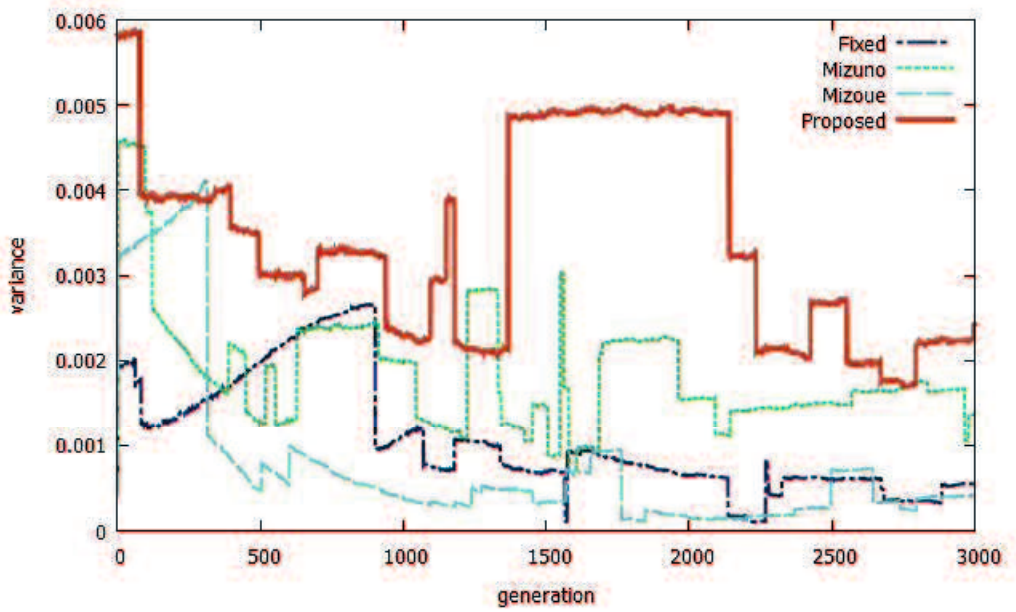


図 5.12: 個体群のばらつきの推移

GA により得られた各従来手法の設定パラメータを表 5.6 に示す。また、同様に提案手法における行列 A を式(5.19)に示す。

表 5.6: GA により得られた各従来法のパラメータ

Method	**Parameter**	**solution**
Fixed	φ	0.68
	γ	0.97
	T	0.88
Mizuno's method	$height_{\varphi}$	0.58
	$slide_{\varphi}$	0.88
	$height_{\gamma}$	0.00
	$slide_{\gamma}$	3.24
	$height_{T}$	24.00
	$slide_{T}$	0.28
Mizoue's mehotd	$weight_{\varphi}$	8.90
	$weight_{\gamma}$	0.70
	$weight_{T}$	0.10

$$A = \begin{bmatrix} P_{stra1/joy} & P_{stra1/ang} & P_{stra1/fear} & P_{stra1/sad} \\ P_{stra2/joy} & P_{stra2/ang} & P_{stra2/fear} & P_{stra2/sad} \\ P_{stra3/joy} & P_{stra3/ang} & P_{stra3/fear} & P_{stra3/sad} \\ P_{stra4/joy} & P_{stra4/ang} & P_{stra4/fear} & P_{stra4/sad} \end{bmatrix} = \begin{bmatrix} 0.60 & 0.35 & 0.06 & 0.02 \\ 0.01 & 0.12 & 0.04 & 0.51 \\ 0.29 & 0.16 & 0.58 & 0.05 \\ 0.09 & 0.35 & 0.31 & 0.41 \end{bmatrix} \quad (5.19)$$

次に、GA の最終世代における優秀個体から得られたパラメータを用いて、再度ターゲットタスクの学習を行った。各手法の平均ステップ数の推移を図 5.13 に示す。このグラフは、1 試行 500 エピソードにおける各エピソードでの終了ステップ数の推移を、30 試行で平均したものである。初期のエピソードでは、エージェントはゴールへ到達できないため、いずれの手法も最大ステップ数である 100 に近い値である。その後、提案手法と Fixed の値が早く減少しており、これは学習のスピードが速いためであると考えられる。300 エピソードでは環境変化が起こり、これまでの最適経路が喪失するため、いずれの手法もステップ数が激増している。その後の再学習が最も早いのは Mizoue であるが、400 エピソード周辺で提案手法が入れ替わり最もステップ数が少なくなっている。

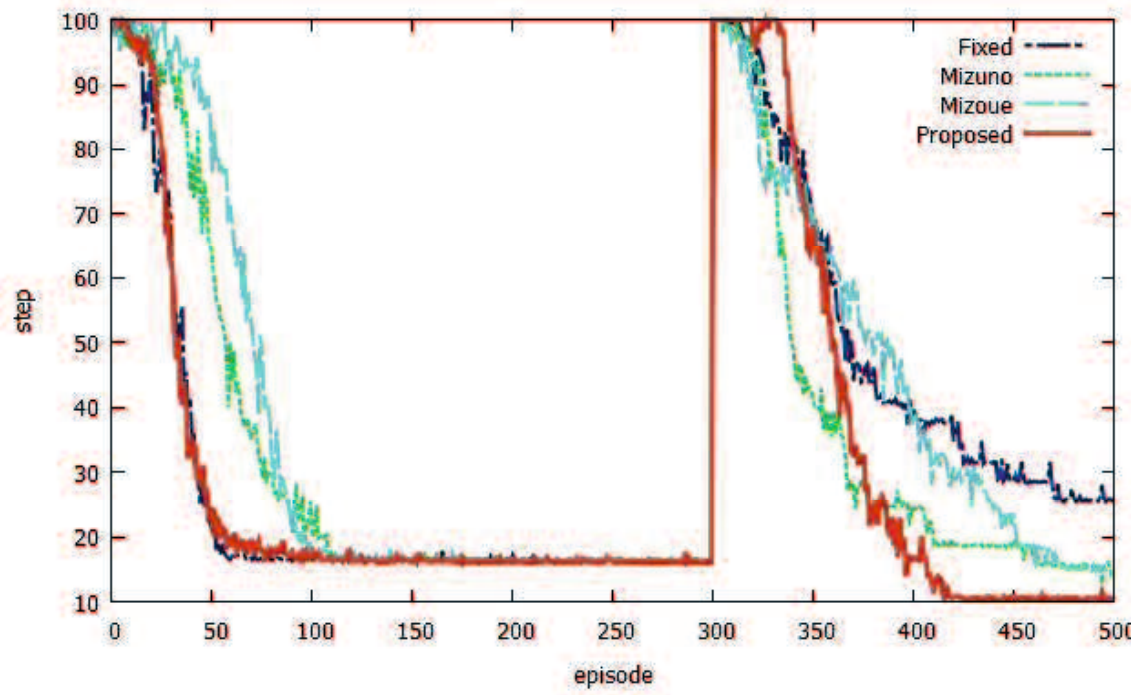


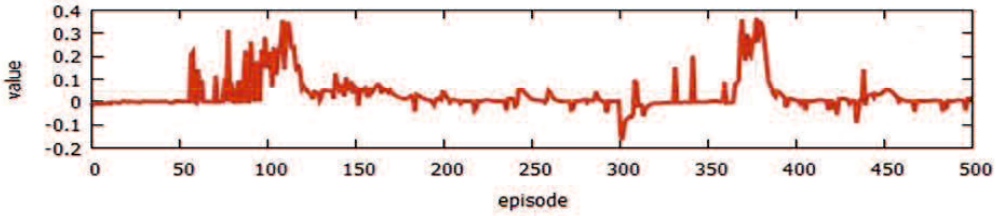
図 5.13: 各手法の平均ステップ数の推移

次に、各手法について、GA の最終世代における優秀個体から得られたパラメータを用いたターゲットタスクの学習における、1 試行中の各メタパラメータの推移を観察する。

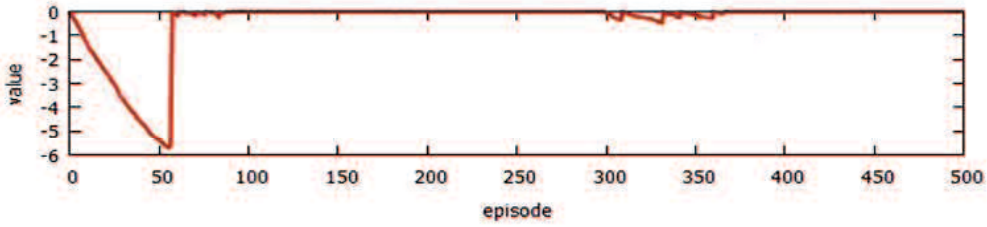
まず、Mizuno 手法の学習における TD 誤差、Down reward、学習係数、割引率、温度定数の推移を図 5.14 に示す。Mizuno 手法は報酬の減少率である Down reward を用いてメタパラメータを変化させる。結果では Down reward の値は初期の学習においては大きく減少しているが、300 エピソードにおける環境変化に対しては値の減少が小さい。また、図から見て取れるように、各パラメータの変動の大きさは Down reward に比例している。

次に、Mizoue 手法の学習における TD 誤差、$TDerror'$、学習係数、割引率、温度定数の推移を図 5.15 に示す。Mizoue 手法は TD 誤差より計算される$TDerror'$ によりメタパラメータを変化させる。本来このモデルは環境変化による TD 誤差の上昇に対してメタパラメータを変更することを目的としているが、本シミュレーションタスクにおいてはゴール報酬が大きく、ゴールに到達しない場合の負の報酬が少ない。そのため、学習が進みゴールに到達できるようになるとメタパラメータが再学習に適切な値に変化し、一方、環境変化時には TD 誤差が小さいためパラメータは学習を抑制する値に変化している。また、図から見て取れるように、Mizuno 手法と同様に各パラメータの変動の大きさは$TDerror'$に比例している。

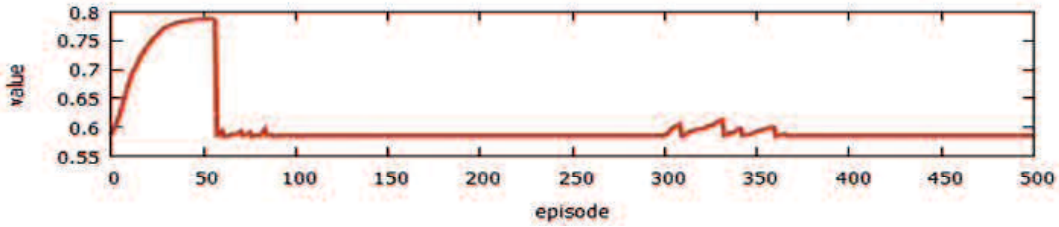
次に、提案システムの学習における $TDerror'$、 *Down reward*、情動状態、学習係数、温度定数の推移を図 5.16 に示す。提案システムにおいては学習係数と温度定数を$TDerror'$および *Down reward* の感覚刺激に応じて情動反応が起こり、情動状態が推移する。その情動状態によってメタパラメータは変更されるが、図から見て取れるように、各メタパラメータの値は感覚刺激に比例しておらず、より複雑なルールが表現されていることがわかる。



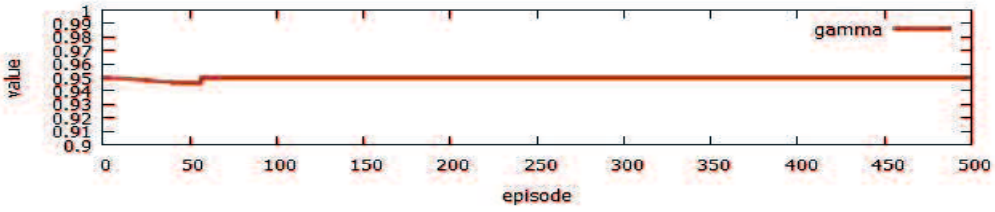
(a) TD 誤差 $TDerror$



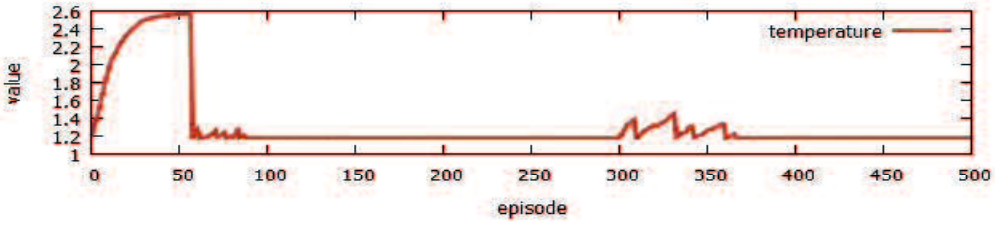
(b) *Down reward*



(c) 学習係数 φ

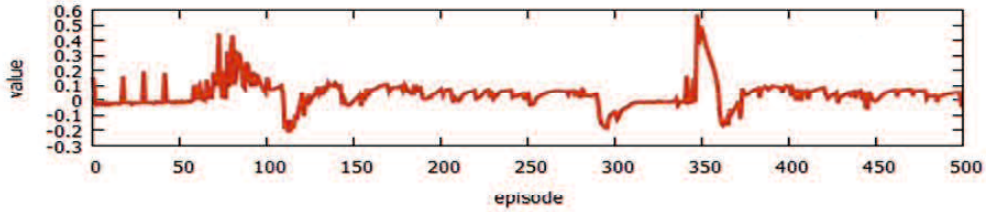


(d) 割引率 γ

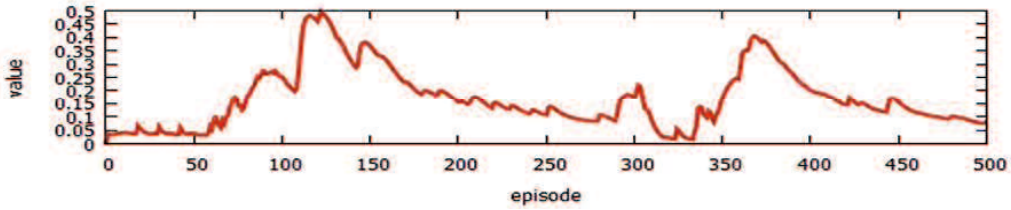


(e) 温度定数 T の推移

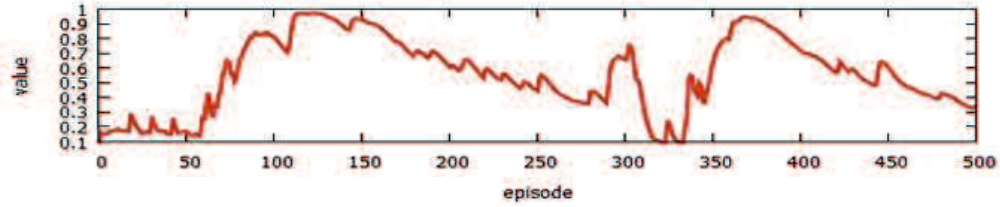
図 5.14: Mizuno における各パラメータの推移



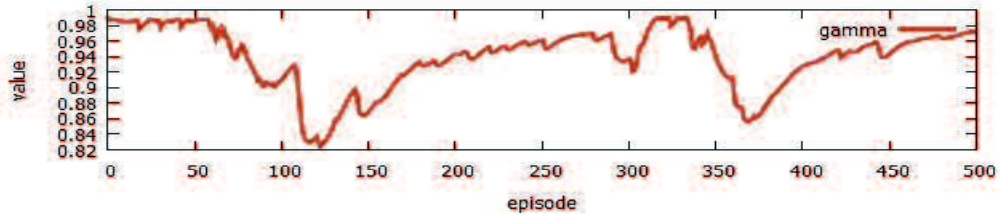
(a) TD 誤差 $TDerror$



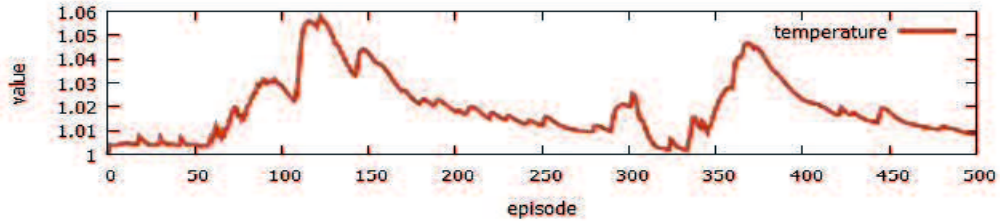
(b) $TDerror'$



(c) 学習係数 φ

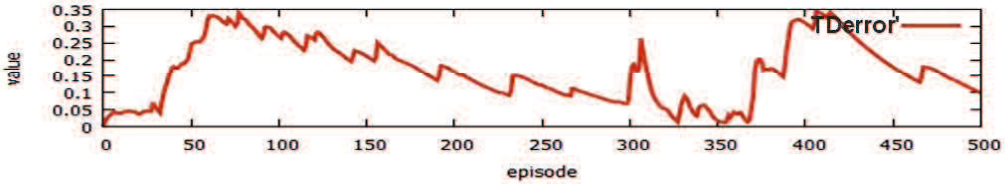


(d) 割引率 γ

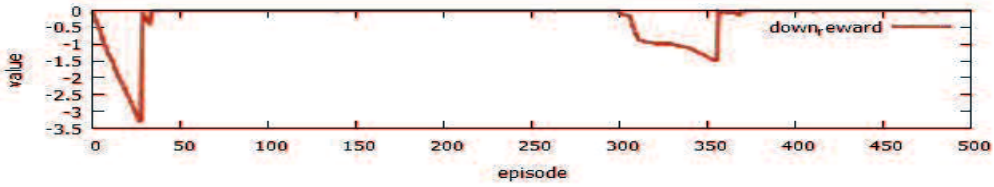


(e) 温度定数 T の推移

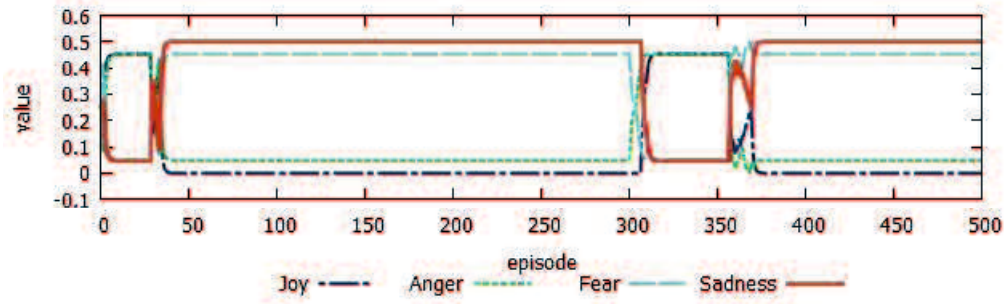
図 5.15: Mizoue における各パラメータの推移



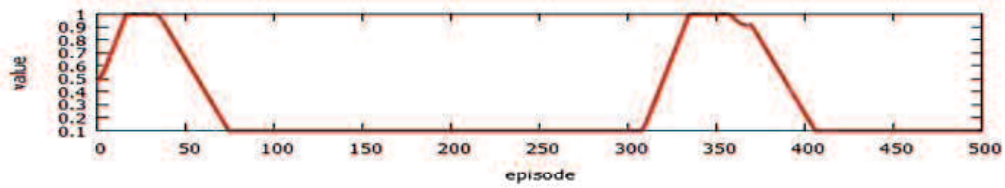
(a) *TDerror'*



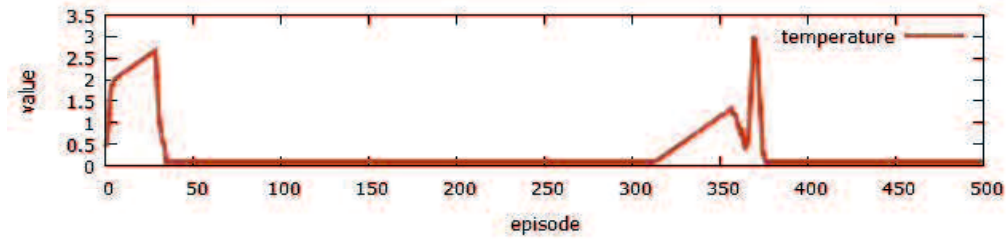
(b) *Down reward*



(c) 情動状態の推移



(d) 学習係数 φ



(f) 温度定数 T の推移

図 5.16: 提案手法における各パラメータの推移

提案システムのメタパラメータ制御をより詳細に観察する。感覚刺激と情動反応の対応を理解しやすくするため、図 5.9 の結果から作成した情動マップを図 5.17 に示す。また、300 エピソードから 400 エピソードまでの情動状態の移り変わりを情動マップ上に示した図を 5.18 に示す。これは、306 エピソード、357 エピソード、365 エピソード、368 エピソードを頂点として簡略化して表示している。提案システムの学習における$TDerror'$、 Down reward、情動状態、学習係数、温度定数の 300 エピソードから 400 エピソードの期間の推移を図 5.19 に示す。説明を容易にするために、306 エピソード、357 エピソード、365 エピソード、368 エピソードに破線を表示している。また、遷移行列 A を結果と比較しやすいように式(5.20)に再び示す。

$$\mathrm{A}=\begin{bmatrix} P_{stra1/joy} & P_{stra1/ang} & P_{stra1/fear} & P_{stra1/sad} \\ P_{stra2/joy} & P_{stra2/ang} & P_{stra2/fear} & P_{stra2/sad} \\ P_{stra3/joy} & P_{stra3/ang} & P_{stra3/fear} & P_{stra3/sad} \\ P_{stra4/joy} & P_{stra4/ang} & P_{stra4/fear} & P_{stra4/sad} \end{bmatrix}=\begin{bmatrix} 0.60 & 0.35 & 0.06 & 0.02 \\ 0.01 & 0.12 & 0.04 & 0.51 \\ 0.29 & 0.16 & 0.58 & 0.05 \\ 0.09 & 0.35 & 0.31 & 0.41 \end{bmatrix} \quad (5.20)$$

まず、300 エピソードの環境の変化により$TDerror'$は大きく上昇し、down reward は小さく減少した。それに対して Anger が誘発され、Fear の情動値は減少している。この時、Sadness の情動値が高いため、φ を減少させる strategy2 と T を減少させる strategy4 の選択確率が高く、φ と T は低い。

306 エピソード（1 番目の破線）を過ぎると学習により$TDerror'$は減少し、壁からの負の報酬が伝搬することで Down reward が減少することにより、Joy と Anger の 2 つの情動値が誘発される。この 2 つの情動は共に φ を増加させる strategy1 の選択確率を高めるため、φ の増加が開始する。また、ともに T を高める strategy3 の選択確率も高めるが、Anger の情動は T を下げる strategy4 の選択確率も高め、打ち消しあうため、T の値の変化は小さい。

357 エピソード（2 番目の破線）になると、Down reward の値が急増することにより Sadness と Fear の情動が急速に誘発され、φ と T は減少する。しかし、365 エピソード（3 番目の破線）では学習が収束することによる $TDerror$ の減少と、Down reward のわずかな減少により、T を高める strategy3 の選択確率を高める Joy と fear の情動が誘発され、T を減少させる strategy4 の選択確率を高める Sadness の情動値が減少することで、Tが急激に増加している。

その後、368 エピソード以降では $TDerror$ の上昇により Sadness が誘発され、φ と T がともに減少している。

このように、感覚刺激に基づき情動反応が起こり、情動行動は学習戦略として効果的なメタパラメータ制御が行われることが示された。

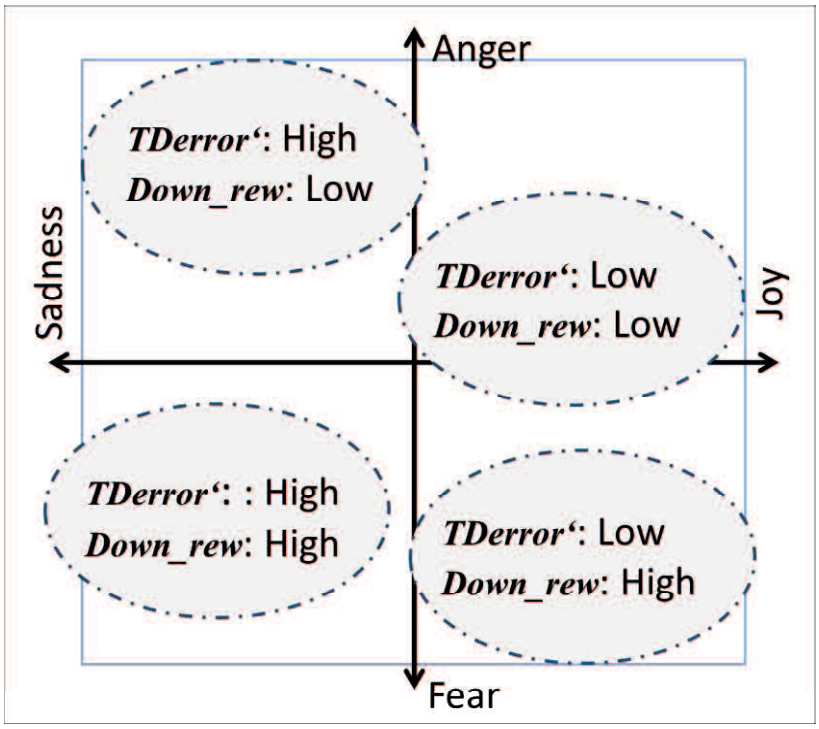


図 5.17: SOM の各感覚刺激の分布

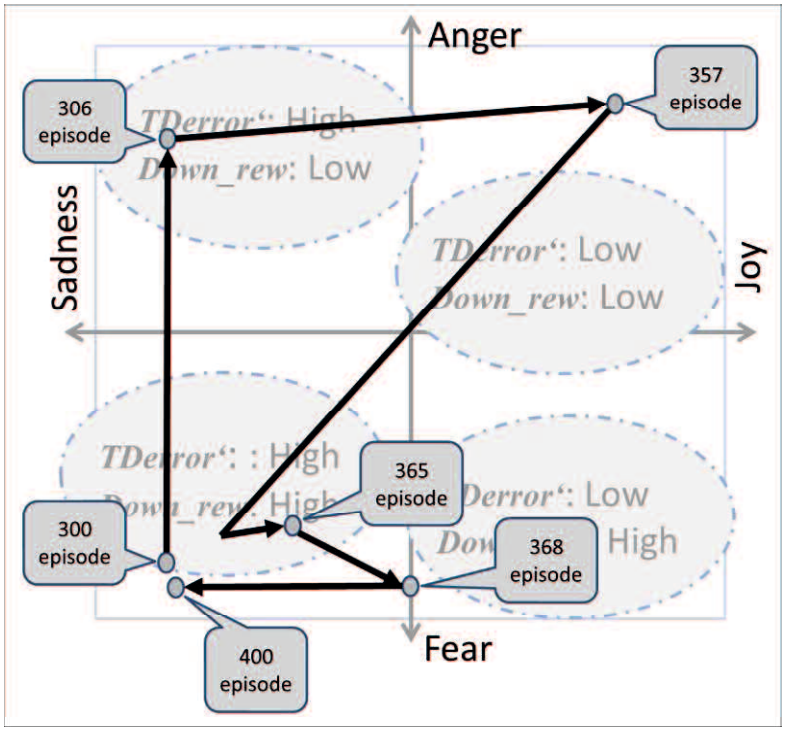
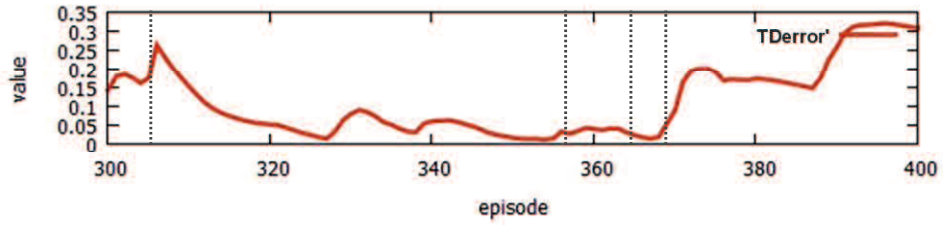
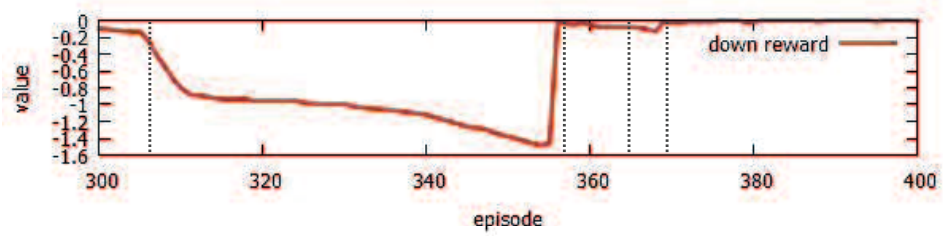


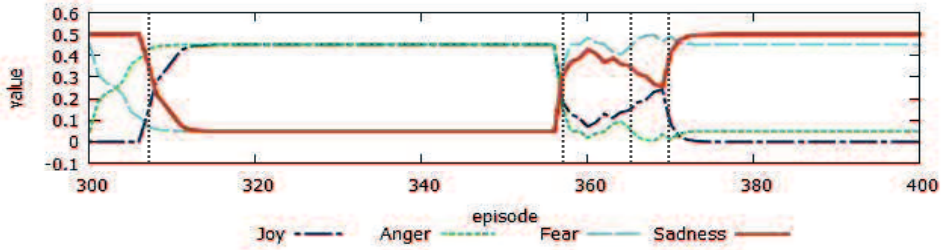
図 5.18: 情動状態の移り変わり



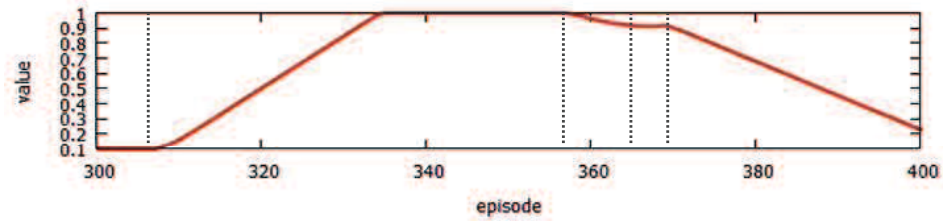
(a) *TDerror'*



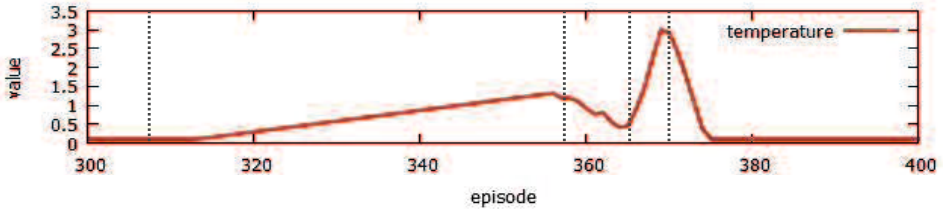
(b)*Down reward*



(c) 情動反応



(d) 学習係数 φ



(f) 温度定数 T の推移

図 5.19: 提案手法における各パラメータの推移

次に、各手法において訓練タスクで獲得した学習戦略およびメタパラメータを用いて、目的タスクの学習を行った。各手法の平均ステップ数の推移を図 5.20 に示す。このグラフは、1 試行 500 エピソードにおける各エピソードでの終了ステップ数の推移を、30 試行で平均したものである。

訓練タスクの学習における初期のエピソードでは、提案手法と Fixed のステップ数が早く減少し、それと比較すると Mizuno 手法と Mizoue 手法は減少が遅かった。一方、目的タスクの学習では、Mizuno 手法と Mizoue 手法は提案手法よりも減少が早い。これは、環境内に壁が多いことが理由として挙げられる。Mizuno 手法と Mizoue 手法は TD 誤差の変化や報酬の減少をもとにメタパラメータ制御を行うため、壁が多い環境のほうがそれらの反応が起こりやすい。200 エピソードで環境変化が起こり、その後、Fixed と Mizuno 手法はゴールに到達できていない。Mizoue 手法と提案手法はステップ数が少しているが、提案手法の方がステップ数の減少が大きい。

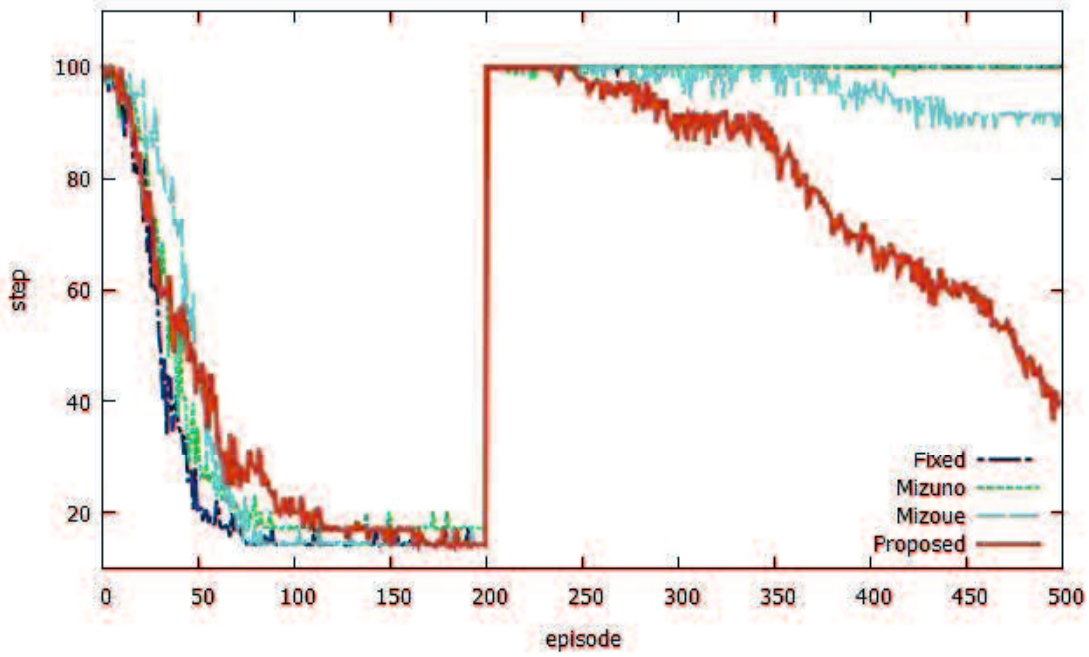


図 5.20: 強化学習を導入した改良型の提案システム

結果より、目的タスクが訓練タスクと異なっていても、訓練タスクの学習で獲得した学習戦略を目的タスクの学習に用いることで、学習を効率化できることが示された。

5.4 考察

本章では、提案システムの強化学習得への応用を提案した。従来の提案システムは、獲得すべき情動行動が目的タスクにおける最適行動を担っていたため、タスクごとに情動形成学習および情動行動学習が必要であった。これらの学習は膨大な繰り返し計算が必要であり、実機ロボットにおいてこれらを行うことは困難となる場合がある。そこで、提案システムを強化学習へ応用し、システムの情動行動は強化学習を効率化する学習戦略として利用されるものとした。学習戦略はタスクに依存しない効率的な学習のためのルールであり、ロボットの目的タスクの行動学習を加速する効果があるが、学習戦略は目的タスクとは異なる訓練タスクを用いて獲得可能である。従って、目的タスクが実機ロボットを用いた行動学習であっても、提案システムは訓練タスクとして目的タスクと類似した条件の計算機シミュレーションにより学習戦略を獲得し、それを実機ロボットの学習に用いることで行動学習を効率化が可能である。本研究では、学習戦略は強化学習のメタパラメータの制御とした。

計算機シミュレーションでは、環境変化が発生する迷路探索問題をタスクとして従来手法と学習の振る舞いを比較した。提案手法および従来手法の各パラメータは実数値 GA により最適化を行った。訓練タスクの学習の結果より、提案手法においては環境変化後の再学習が適切に行えることが示された。学習中のメタパラメータの推移を観察すると、従来手法においてはメタパラメータの決定に用いる因子*TDerror'*や*Down reward*に比例しており、シンプルなルールで表現されている。一方、提案手法ではメタパラメータとその決定に用いる因子が比例関係でなく、複雑なルールによってメタパラメータが制御されている。これらのルールは試行によって獲得された Cognition モジュールや遷移行列 A の各パラメータを解析することで理解することができる。

また、訓練タスクの学習によって獲得された各手法のパラメータを用いて、目的タスクの学習を行った。目的タスクは訓練タスクと構造が大きく異なり、従来手法においては訓練タスクによって獲得したパラメータを用いても再学習が適切に行えなかった。しかし、提案手法においては目的タスクにおいても再学習能力が従来手法に比べて高いことが示された。

第6章　　結言

より高度で人間らしいロボットシステムの開発のために生物がもつ情動をモデル化し、その工学的応用を試みる研究が数多く存在するが、それらの研究のほとんどは人間が進化過程で既に獲得した情動の機能をモデル化している。しかし、人間が情動発生のメカニズムを進化の過程や後天的な経験により獲得したように、近年注目されている自律ロボット開発の観点においては人工情動も環境に適応的なシステムの構成方法が必要である。そこで本研究では、人間の情動反応を模倣させるのではなく、必要な情動反応をロボット自身が外部環境に適応して形成することが可能なシステムとして、マルコフ情動モデルに基づく自律ロボットの意思決定システムを提案した。

まず、第 3 章では、提案するマルコフ情動モデルに基づく自律ロボットの意思決定システムについて述べた。提案システムは、感覚刺激の認識から情動反応、そして情動行動が生成されるまでの一連の処理を実装するシステムの枠組みである。提案システムは、これらの処理を実装する上で必要な情動反応や情動行動のルールを人間の情動に基づき事前に手動設計する必要はなく、システムを実装するロボットのスペックや対象とするタスクに適応して自動設計する機能を有する。システムは次に述べる 2 つの情動学習過程により情動反応および情動行動に関するルールを構築する。1 つ目の情動形成学習は感覚刺激と情動反応の対応付けを行う学習過程である。この学習は自己組織化写像（SOM）を用いた感覚刺激のクラスタリングによって行われる。2 つ目の情動行動学習は情動反応と情動行動を関連付ける学習過程である。この学習はタスクの試行におけるシステムの設計パラメータを適切に調節することにより行われる。また、提案システムにおける情動反応や情動行動のルールの記述に関して、情動という媒体を介することでより複雑な処理を表現可能であり、さらに、獲得されたそれらのルールをヒトが意味的に理解すること容易である。システムに関する評価シミュレーションの結果では、2 つの情動学習過程により適切な行動決定法が自動構築され、有効な行動決定が生成されることが確認された。しかし、提案システムはその情動形成学習に SOM を用いず試行錯誤により手動設計したシステムに比べて、獲得されたロボットの行動決定法の性能が劣る評価となった。この原因は、感覚刺激の教師無し学習時に、感覚刺激の発生確率を考慮しなかったため、SOM の競合層上に発生頻度の低い無駄な状態領域が構築されたためだと考えられる。

そこで、第 4 章では、SOM の競合層上における無駄な状態空間を減少させることを目的に、ロボットの経験に基づく情動の再形成学習を導入した。感覚刺激の予測値に基づく学習サンプルのみならず、ロボットがタスク中に得た感覚刺激をオンラインで学習可能とな

るようにシステムの部分的な改良を行った。改良手法の評価シミュレーションの結果より、情動の再形成学習により SOM 中の不必要な感覚刺激への対応付けが減少し、生成されるロボットの行動決定法の性能を向上させることが示された。

ここまでの研究では、システムにより設計される行動決定法は学習する行動タスクに依存するため、目的とする行動タスク毎に学習する必要があった。システムにおける学習過程は膨大な繰り返し計算が必要であるため、この問題点は実機ロボットにおける学習を困難とする場合があると考えられる。そこで、第 5 章では、提案システムのさらなる汎用性を示すために、提案システムの強化学習への応用を行った。ここで、提案システムにおける情動は学習を効率的に行うための学習戦略を提供する。学習戦略はタスクに依存せず、あらゆるタスクに再利用可能な知識の効果的な利用方法を指す。訓練タスクを用いた人工情動の事前学習により適切な学習戦略を発見し、得られた学習戦略を利用することで目的タスクの学習を効率化する。本研究では、学習戦略は具体的に強化学習のメタパラメータの適応的な調節とした。環境が動的に変化する迷路探索タスクに関する評価シミュレーションにより従来手法と比較し、提案手法が従来手法よりも複雑なルールを表現でき、それらを自動的に生成できることを示した。これらは実際のロボットへの対応におけるシステムのより汎用性の高い利用方法であると考えられる。例えば、目的タスクに類似したタスクのシミュレーションにおいて学習戦略を獲得し、その学習戦略を実ロボットによる行動学習に利用することで、実ロボットにおけるオンライン学習を効率化することができると考えられる。

本研究の大きな目的でもある自律ロボット開発の観点では、ロボット個体による個性や、プログラミングされた動作以外の行動ルールを自律的に獲得することは興味深いと考えられている。しかし、実際のロボットの制御を行う機構には適切な安全設計を行う必要がある。生物において脳機能は進化により認識処理過程を計算的処理過程と短絡的処理過程に分化し、後者の短絡的処理過程が情動を指す。短絡的処理過程である情動は、特定の刺激に対して状況の推測や情報分析を経ずに行動をある方向へと駆り立てる認識処理過程である。しかし、人間は短絡的処理過程の結果から直接行動決定を行っているわけではなく、その結果を感情として感じ取り、外部の状況とそれらを統合的に情報処理して行動決定を行う。つまり、人間が欲のままに行動するのではなく社会的適応することと同様に、ロボットにおいても情動から行動への変換の際には適切な行動決定器を介する必要がある。将来的にシステムをロボットへ実装する際には、提案システムは行動の多様性を高めるために補助的に使用し、実際のロボットの制御を行う機構には適切な安全設計を行うことが良

案である。また、5 章で述べた強化学習への応用のように行動学習の効率向上を目的として用いることも有効であると考えられる。

本研究においては、結果の可読性を高める理由から感覚刺激の入力数、および、基本情動の数を小さく設定し、Cognition モジュールはシンプルな 2 次元 SOM によって構成した。しかし、近年 IoT（Internet of Things）という言葉が一般的になってきたように、インターネットの普及により社会に存在するあらゆるモノにインターネットが接続される時代が到来しようとしている。その時、ロボットの感覚刺激は自身が搭載するセンサーやカメラだけでなく、他ロボットのカメラや環境設置型のセンサーなど膨大な数の感覚装置にアクセスが可能となる。感覚刺激の数が増えると、2 次元 SOM では表現能力が不足し、適切な情動の形成が困難となる。本研究の発展として、Cognition モジュールを更に高次な状態分類器により構成し、基本情動の数を増加させることにより、さらに複雑な情動形成が可能と考えられる。

また、本研究では基本情動の名称として Joy、Anger、Fear、Sadness を用いた。しかしながら本研究においては呼称の役割しか担わず、意味的には人間の基本情動とは対応しない。これらは、本研究の人間の情動をモデル化するのではなく、ロボット自身が情動の獲得を行うことを目的としていることに起因する。「怒り」の情動は昆虫に至る多くの生物に観測される感情状態であるが、「愛」の情動は人間のような高等動物のみが獲得していると考えられる。また、心理学者 Maslow は欲求階層説において、人間の欲求は最下位層から順に生理的欲求、安全欲求、親和欲求、自我欲求、そして最上位層の自己実現欲求までの 5 段階に分類でき、低次の欲求が満たされると高次の欲求を満たすように動機づけられると提唱している。これらの知見から、人間は進化によって上位層の欲求および、それに伴う高度な情動を獲得したと考えられる。人間がまだ獲得していない、新たな機能性の情動をロボットが自律的に獲得できることはロボットの進化であり、非常に興味深いと考えている。

参考文献

[1] 久保田直行 , 脇坂史帆 , 小嶋宏幸 , “情動モデルを用いたパートナーロボットに関する研究 : 仮想現実空間の構築と人間との相互作用” , 知能と情報（日本知能情報ファジィ学会誌）Vol.20, No.4, pp.449-460, 2008.

[2] 大林正直, 呉本尭, 小林邦和, "インテリジェントコンピューティング", 山口大学工学部知能情報工学科 生体情報システム工学研究室, 2010

[3] 鈴木薫, 金澤博史, “感情動因学習モデルを用いたペットロボット”, 東芝レビュー, Vol.56, No.9, 2001.

[4] ヤン スンハ, 安藤繁, “高分解オプティカルフローに基づく表情推定 : 顔表面の並進移動と収縮テンソルのモデル化と同時検出”, 電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解 111(379), 297-300, 2012.

[5] 小林宏, 原文雄, “ニューラルネットワークによるヒトの基本表情認識”, 計測自動学会論文集, Vol.29, No.1, pp.112-118, 1993.

[6] 川上文雄, 山田寛, 原島博, 森島繁生, “3 次元感情モデルに基づく表情分析・合成システムの構築”, 電子情報通信学会技術研究報告. HCS, ヒューマンコミュニケーション基礎 95(552), 7-14, 1996.

[7] Meng-Ju Han, Chia-How Lin and Kai-Tai Song, “Robotic Emotional Expression Generation Based on Mood Transition and Personality Model”, IEEE TRANSACTIONS ON CYBERNETICS, Vol.43, No.4, pp.1290-1303, 2013.

[8] 伊藤加寿子, 三輪洋靖, 忽滑谷裕子, 齊藤稔, Massimiliano Zecca, 高信英明, Stefano Roccalla, Maria Chiara Carrozza, Paolo Dario, 高西淳夫, “ヒューマノイドロボットと人間とのインタラクションにおけるロボット評価システムの開発”, ロボティクス・メカトロニクス講演会’06 講演論文集, 2P1-A15, 2006.

[9] J. Moren, C. Balkenius, “A Computational Model of Emotional Learning in the Amygdala”, Cybernetics and Systems 32(6), pp.611-636 , 2001.

[10] M. Jamali, M. Dehyadegari, A. Arami, C. Lucas, and Z. Navabi, “Real-time embedded emotional controller”, Neural Computing & Applications, Vol.19, No.1, pp.13-19, 2009.

[11] A. Arami, C. Lucas, and M. Nili-Ahmadabadi, "Attention to Multiple Local Critics in Decision Making and Control", Expert Systems with Applications, Vol.37, pp.6931-6941, 2010.

[12] H. Rouhani, M. Jalili, B. N. Araabi, W. Eppler and C. Lucas: “Brain Emotional Learning Based Intelligent Controller Applied to Neurofuzzy Model of Micro Heat Exchanger”, Expert Systems with Applications, Vol.32, No.3, pp.911-918, 2007.

[13] T. Kuremoto, T. Ohta, K. Kobayashi, and M. Obayashi, “A Dynamic Associative Memory System Adopting Amygdala Model”, Artificial Life and Robotics, Vol.13, No.2, pp.478-482, 2008.

[14] Ehsan Lotfi, A. Keshavarz, “A simple Mathematical Fuzzy Model of Brain Emotional Learning to Predict Kp Geomagnetic Index”, International Journal of Intelligent System and Applications in Engineering, Vol.2 No.2, pp.22-25, 2013.

[15] 宅野雄大，大林正直，呉本尭，小林邦和, ”情動モデル融合型強化学習システム”，平成 22 年電気学会電子・情報・システム部門大会講演論文集, pp.126-131, 2010.

[16] M. Obayashi, T. Takuno, T. Kuremoto, K. Kobayashi, “An Emotional Model Embedded Reinforcement learning System”, Proc. of IEEE International Conference on Systems, Man, and Cybernetics, (SMC2012), pp.1058-1063, 2012.

[17] Francois Michaud, “EMIB-Computational Architecture Based on Emotion and Motivation for Intentional Selection and Configuration of Behavior-Producing Modules”, Cognitive Science Quarterly, 2002,

[18] Sajal Chandra Banik, Keigo Watanabe, Kiyotaka Izumi, “Improvement of group performance of job distributed mobile robots by an emotionally biased control system”, Artif Life Robotics(12) , pp.245-249 ,2008

[19] Sajal Chandra Banik, Keigo Watanabe, Kiyotaka Izumi, "Generation of Cooperative Behavior of Robots Using a Fuzzy-Markov Emotional Model", International Journal of System Signal Control and Engineering Application 1(1):101-109, 2008.

[20] MENG Qing-mei, WU Wei-guo, “Artificial model based on finite state machine”, J. Cent. South Univ. Technol. No.15, pp694-699, 2008.

[21] Even Daglarli, Hakan Tameltas, Murat Yesiloglu, “Behavioral task processing for cognitive robots using artificial emotions”, Neurocomputing 72 , pp.2835-2844 , 2009

[22] M. Chandra, Analytical Study of A Control Algorithm Based on Emotional Processing, M.S. Dissertation, Indian Institute of Technology Kanpur, 2005.

[23] Christopher P. Lee-Johnson and Dale A. Carnegie, “Mobile Robot Navigation Modulated by Artificial Emotions”, IEEE Transactions, Man, and Cybernetics, Part B: Cybernetics, vol.40, No.3, pp.469-480, 2010.

[24] SHI Xue-fei, WANG Zhi-liang, PING An, ZHANG Li-kun, “Artificial emotion model based on reinforcement learning mechanism of neural network”, The Journal of China Universities of Posts and Telecommunications, 18(3), pp.205-209, 2011.

[25] Ho Seok Ahn, “Designing of a Personality Based Emotional Decision Model for Generating Various Emotional Behavior of Social Robots”, Advances in Human-Computer Interaction”, Vol.2014, 14 pages, 2014.

[26] Qing Zhang, Sungmoon Jeong, Minho Lee, “Autonomous emotion development using incremental modified adaptive neuro-fuzzy inference system”, Neurocomputing 86, pp.33-44, 2012.

[27] 大田智範, 呉本尭, 小林邦和, 大林正直, ”情動-連想 記憶システム”, 第 15 回計測自動制御学会中国支部 学術講演会論文集, pp.62-63, 2006.

[28] T. Kuremoto, M. Obayashi, K. Kobayashi, “A Chaotic Memory System Accelerated by an Emotional Model”, In Insights into Amygdala: Structure, Functions and Implication for Disorders, pp.229-254, 2012.

[29] 宇都俊佑, 大林正直, 呉本 尭, 間普真吾, 小林邦和, ”色 彩特徴による情動を考慮した強化学習システム”, 第 22 回計測自動制御学会中国支部学術講演会論文集, pp.90-91, 2013.

[30] T. Kuremoto, T. Tsurusaki, K. Kobayashi and M. Obayashi, “A Model of Emotional Intelligent Agent for Cooperative Goal Exploration”, Intelligent Computing Theories, Lecture Note in Computer Science, Vol. 7995, pp. 21-30, 2013.

[31] Kuremoto, T., Obayashi, M., Kobayashi, K., and Feng, L, “Autonomic Behaviors of Swarm Robots Driven by Emotion and Curiosity”, Life System Modeling and Intelligent Computing, Lecture Notes in Bioinformatics, Vol. 6330, pp.541-547, 2010.

[32] Takashi Kuremoto, Masanao Obayashi, Kunikazu Kobayashi, Liang-Bing Feng, “An Improved Internal Model of Autonomous Robots by Psychological Approach”, Cognitice Computation, Vol. 3, No.4, pp.501-509, 2011.

[33] 堀哲郎, “脳と情動ー感情のメカニズム”, 共立出版株式会社, 1991.

[34] 大山良樹, 林田一志, 行待寿紀, “ストレスと情動反応について”, 明治鍼灸医学 第 1 号, pp.91-97, 1992.

[35] E. T. Rolls, (eds. Y. Oomura), “Emotions”, Japan Sci. Soc. Press/Karger, pp.325-344, 1986

[36] ジョセフ・ルドゥー, “エモーショナル・ブレイン - 情動の脳科学”, 東京大学出版会, 2003.

[37] Russell, James A, “A circumplex model of affect”, Journal of Personality and Social Psychology, Vol.39(6), pp.1161-1178, 1980.

[38] McDougall, “Review of An Introduction to Social Psychology” , Psychological Bulletin, Vol.5, pp.385-91, 1908.

[39] Watson, “Behaviorism (2nd edition)”, New York Norton, 1930.

[40] Arnold, “Emotion and personality”, Vol.1-2, New York Columbia University, 1960.

[41] Plutchik, “Emotion a Psycho-Evolutionary Synthesis”, New York Harper, 1980.

[42] Panksepp, “Toward general psychobiological theory of emotions”, Behavioral and Brain sciences, Vol.5, pp.407-468, 1982.

[43] Tomkins, “Affect,imagery and consciousness”, New York Springer, Vol.1, 1962.

[44] Weiner and Graham, “From an attributional theory of emotion to developmental psychology”, A round-trip ticket Social Cognition, Vol.4, pp.153-179, 1986.

[45] E. T. Rolls, “A theory of emotion, and its application to understanding the neural basis of emotion, Neural and ChemicalControl” ,Japan Scientific Societies Press, 1986.

[46] Johnson-Laird and Oatley, “Towards a cognitive theory of the emotions”, Cognition and Emotion, Vol.1, pp.29-50, 1987.

[47] Gray, “The neurophsychology of anxiety”, Oxford University Press, New York, 1982.

[48] Izard, “Human Emotions”, New York Plenum Press, 1977.

[49] Ekman, “Emotions in the Human Face”,　London Cambridge University Press, 1982.

[50] Satoshi Kurihara, Shigemi Aoyagi, Rikio Onai, Toshiharu Sugawara, “Adaptive selection of reactive/deliberate planning for a dynamic environment”, Robotics and Autnomous Systems 24 (1998), 183-195, 1998.

[51]K. Doya, G. Hatano, N. Tanabe, “Metalearning, neuromodulation, and emotion”, Affective Minds, Elsever Science, pp.101-104, 2000.

[52]K. Doya, “Complementary roles of basal ganglia and cerebellum in learning and motor control”, Current Opinion in Neurobiology, Vol.10, No.6, pp.732-739, 2000.

[53]銅谷賢治, 石井信, “学習ダイナミクスの制御と脳の物質機構”, システム/制御/情報 : システム制御情報学会誌, 50(8), pp.303-308, 2006.

[54]水野 純也, 村越 一支, “神経修飾物質に対応付けた強化学習パラメータの制御法”, 電子情報通信学会, 信学技報 NC2002-102, pp.83-88, 2002.

[55] Kobayashi, K., Mizoue, H., Kuremoto, T., and Obayashi, M., A Meta-learning Method Based on Temporal Difference Error, Lecture Notes in Computer Science (LNCS), Vol.5863, pp.530-537, 2009.

[56]溝上裕之, 小林邦和, 呉本尭, 大林正直, “TD 誤差に基づく強化学習のメタパラメータ学習法”, 電気学会論文誌 C, 129 巻,9 号, pp.1730-1736, 2009.

[57] 秋口俊輔, 前田陽一郎, ”目標選択型 Q-Learning を用いた自律移動ロボットの情動行動学習”, 第 22 回ファジィシステムシンポジウム論文講演集, 7E3-2, pp.609-614, 2006.

[58] D.E. Rumelhart, G.E. Hinton and R.j. Williams, “Learning Representations by Back Propagation Error”, Nature, vol. 323-9, pp. 533-536, 1986.

[59] T.コホネン, “自己組織化マップ”, シュプリンガー・フェアラーク東京出版, 1996.

[60] Sebastian Thrun, Wolfram Burgard, Dieter Fox 著 上田隆一 訳 “確率ロボティクス” ,株式会社　毎日コミュニケーションズ, 2007.

[61] Thomas Kollar and Nicholas Roy, “Trajectory Optimization using Reinforcement Learning for Map Exploration”, The International Journal of Robotics Research, 27, pp.175-195.

[62] Peng Li, Xinhan Huang, Shengyong Wang and Jean Dezert, “SLAM and Path Planning of Mobile Robot Using DSmT”, Journal of Software Engineering 7 (2), pp.46-67, 2013.

[63] Ali Marjovi and Lino Marques, “Optimal Swarm Formation for Odor Plume Finding”, IEEE Transactions on Cybernetics, Vol.44, No.44, No.12, 2014.

[64] Ali Marjovi, Lino Marques, “Optimal spatial formation of swarm robotic gas sensors”, Auton Robot, 35, pp.93-109, 2013.

[65] David Portugal, Rui P. Rocha, “Distributed multi-robot patrol: A scalable and fault-tolerant framework”, Robotics a and Autonomous Systems, 61, pp.1572-1587.

[66] Ali Marjovi, Lino Marques, “Multi-robot olfactory search in structured environments” , Robotics and Autonomous Systems 59, pp.867-881, 2011

[67]Ziyuan Liu, Georg von Wichert, “Extracting semantic indoor maps from occupancy grids”, Robotics and Autonomous Systems, 62, 6663-674.

[68]荒木天外, 竹村憲太郎, 怡土 順一, 松本 吉央, 高松 淳, 小笠原 司, “汎用三次元環境地図を用いた移動ロボットナビゲーションのための地図生成”, 日本ロボット学会誌, Vol.28, No.1, pp.106-111, 2010.

[69] Miguel Julia, Arturo Gil and Oscar Reinoso, “Searching Dynamic Agents with a Team of Mobile Robots”, Sensors 12, pp.8815-883i, 2012.

[70] F. Mondada, M. Bonani, X. Raemy, J. Pugh, C. Cianci, A. Klaptocz, S. Magnenat, J.-C. Zufferey, D. Floreano, and A. Martinoli: “The e-puck, a robot designed for education in engineering”, In Proceedings of the 9th Conference on AutonomousRobot Systems and Competitions, Vol.1, p59–65(2009).

[71] “Ecole Polytechnique Fedrerale Industrial Relations Office”, http://www.e-puck.org

[72] “Webots robot simulator”, http://www.cyberbotics.com/