目 次

1　序論

本章では，まず第 1 節にて，これまでのロボットの自動制御と学習について述べる．次に，第 2 節では，ロボットにとってのセンサ情報の役割について述べる．第 3 節では，近年，多くのセンサを搭載したマルチタスク・ロボットが現れ始めていることについて述べる．第 4 節では，これまでは人間による予測にって，ロボットに搭載するセンサが選定されていたことについて述べる．そして，今後はそれが難しい問題になることについて述べる．第 5 節では，近年ではロボットに学習を適用する事例が多くあるが，学習ロボットのシステムに多くのセンサ情報を入力することは，決して望ましくないことであることについて述べる．

以上までの節で述べたことを踏まえて，第 6 節では，その問題を解決するための本研究の目的を述べる．第 7 節では従来研究について述べ，第 8 節では本研究のアプローチについて述べる．最後に，第 9 節では，以降の本論文の構成について述べる．

1.1　ロボットの自動制御と学習

これまでのロボットは，自動制御を適用されたものだった [1]．すなわち，人間であるプログラマが，プログラミングした通りの動作しかできないものであった．このロボットの動作は，ある状況のときにこの動作をするというように—条件分岐的に—生成されるものである．すなわち，自動制御のロボットがタスクを実行するとき，そのロボットが直面する全ての状況が，人間の予測通りであることが前提であった．

しかしながら，プログラマがタスク・プログラミングすることはとても困難である．なぜなら，タスク・プログラミングするときに，ロボットが直面する全ての状況を予測しなければならないからである．さもなければ，ロボットはプログラマが意図しない動作をする恐れがある．例えば,「マグカップにインスタントコーヒーを淹れて人間に届ける」というタスクをロボットに教えることを考える．これを実行するには，およそ次の手順で果たされるだろう．ロボットはまず，コーヒーの粉が入った袋をとる．次に，マグカップを取り出してテーブルに置き，袋の中にあるコーヒーの粉をマグカップに移す．そして，ポットの中にあるお湯を，ある量だけマグカップに注ぐ．最後に，コーヒーを頼んだ人間のもとにそれを届ける．人間がこれらの手順を教えるとき，マグカップが倒れているかもしれないこと，マグカップがいつもと違う棚の中にしまってあるかもしれないこと，その袋の中のコーヒーの粉がないかもしれないことなど，ありとあらゆる予測をしなければならない．さもなければ，ロボットは倒れたままのマグカップの上にお湯を注いだり，延々とマグカップを探したり，コーヒーの粉を入れないままお湯を注いだりするかもしれないのである．

このように，人間であるプログラマが自動制御のロボットをタスク・プログラミングするとき，そのロボットが直面する状況を可能な限り予測しなければならない．仮にそれができたとしても，ロボットは有限の情報処理能力しかないため，無限に近い場合の数だけ考えられる状況の中から，その対処を選び出すことができない．いわゆる，フレーム問題[1] である．したがって，多くの研究者が，従来の自動制御を適用したロボットにタスクを実行させるのでは限界があるとしている．

近年では，このような自動制御に替わるもとして，多くの研究者がロボットの学習に関心を集めている．日本工業標準調査会によれば,「学習制御」という言葉を以下の通りに定義している [25].

> 過去の時点で得られた制御過程をもとに，制御パラメータ，及び／又はアルゴリズムを，逐次所要の条件を満たすように修正していく制御.

多くの研究者が，ロボットの学習としてソフト・コンピューティングについて研究している．ソフト・コンピューティングの代表的なものには，ニューラル・ネットワーク，遺伝的アルゴリズム，強化学習などがあり，それぞれに学習のモデルと固有の特長がある．ニューラル・ネットワークは，脳細胞の構造とその働きを模した学習方法である．ニューラル・ネットワークは，幾つかの例題とそれらの解答を事前に学習することで，未知の問題に対する解答も推論的に導き出せる[2]．遺伝的アルゴリズムは，生物の突然変異・交叉・淘汰に倣った学習方法である．遺伝的アルゴリズムは，膨大な解の候補の中から，現実的な時間内で準最適解を求めることができる．これらの学習は，それぞれの特長を活かして，互いに組み合わせて使われることもある．強化学習については第 3 章で後述するため，ここでの説明は割愛する．

これらの学習は，ナップザック問題，巡回セールスマン問題などのようなNP-完全問題[3] でも，膨大な解の候補の中から，現実的な時間内で極めて良好な解を求めることができると知られている．ただし，学習によって求められた解は，必ずしも最適なものであるとは限らない．すなわち，その最適性は保証されていないものである．

このように，これらの学習の方法は，多くの成果に基づく事実によって研究者の間で期待され，今日でも盛んに研究されている．ただし，今日までに

---

[1] フレーム問題は，John McCarthy 氏と Patrick J. Hayes 氏によって初めて言及されたものである．フレーム問題の有名な例としては，Daniel Dennett 氏が Cognitive Wheels : The Frame Problem of AI の中で述べた「バッテリーと時限爆弾」が挙げられる．

[2] 推論とは，論理学における厳密な意味では，前提—命題の集合—から結論—新たな命題—を導き出すことである [23]．しばしば，ニューラル・ネットワークが未知の問題に対する解を出力することを「推論する」というが，決してこのような意味ではない．したがって，厳密な意味と画するために，ここでは「推論的」と表現した．

[3] NP-完全という概念は，S.A.Cook 氏によって提唱された概念である．クラス NP—多項式時間アルゴリズムが発見されておらず，現実的な時間内で解を求めることができない問題の分類—に含まれるある問題が NP-完全であるとは，その問題がクラス NP のなかでも最も難しいということである [22].

研究されているロボットの学習が，必ずしも最善な方法であるとは限らない．Rodney A. Brooks 氏が開発した六足ロボット Genghis [7] のシステムのように，学習に依らずとも，ロボットを十分に制御できる可能性も残されている．このことから，ロボットの学習が主流に至るには，今しばらくの時間を要するようである．

1.2 ロボットとセンサ情報

近年では，多彩で高性能なセンサが多く開発されている [5]．最近では，解析機能や情報処理機能を有するもの—スマート・センサ，またはインテリジェント・センサと呼ばれるもの— まで開発されている [20]．また，そのようなセンサを搭載したロボットのためのシステムも多く研究されている [6] ．

センサが計測した物理量は，ほとんど多くの場合，電気的な量—電圧，電流—で表現される信号で出力される．この電気信号をロボットやコンピュータに入力するとき，その電気信号は更に A/D コンバータによって量子化される．すなわち，連続量である電気信号から離散量であるビット値に変換される．本論文では，このビット値のことを指して，以下では「センサ情報」と呼ぶことにする．

さて，ロボットに適用されるものが自動制御にせよ学習にせよ，そのシステムには入力系が存在する．ロボットがあるタスクを実行するとき，多くの場合，適応的な振る舞いが求められるからである．そのため，システムの入力にはセンサ情報が用いられる．そのシステムが出力する制御量をセンサによって計測してフィードバックしなければ，そのシステムに再び入力するべき目標量を決定できないからである．すなわち，システムが適応的な動作を出力できないのである．

ロボットが自分の腕を上げることを例に挙げる（図 1）．このとき，ロボッ

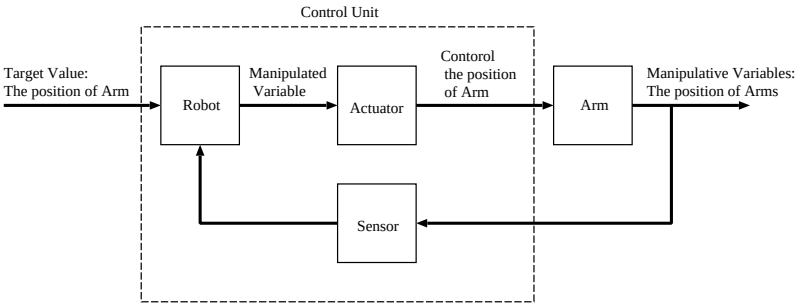


図 1: 腕を上げるときのフィードバックによる適応的な出力（ブロック線図）

トは自分の腕をどれだけ上げればよいのかを確認するために，自分の腕の位置を知る必要がある．ここで，もしロボットが自分の腕の現在の位置が分か

らなければ，垂直な位置までの偏差を求めることができない．すなわち，制御系に入力する目標値を決定することができないのである．それどころか，自分の腕が垂直に上がったとしても，人間が何らかの手段で教えない限り，ロボットは自分自身でそれに気付くことはできないのである．これでは，ロボットはいつまで経ってもタスクの実行を完了できないことがわかる．

したがって，制御対象がロボットであるシステムには，必ず入力系が存在しなければならない．そして，その入力系にはセンサが対応する．一般に，同じ物理量を計測するセンサの個数が増えれば，計測誤差を是正することができる．また，異なる物理量を計測するセンサの個数が増えれば，その値に応じて出力値を適切に調整することができる．そのため，ロボットはより多くのセンサを搭載することにより，より多くの環境に適応できるのである．また，より多くのセンサを搭載することにより，より多くのタスクを実行できるようになるのである．

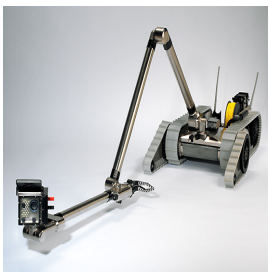
1.3 マルチタスク・ロボットとセンサ

今日では，多彩で高性能なセンサの開発によって，ロボットはより多くの環境に適応できるようになった．これにより，多くのロボットが人間社会における様々な場所や場面で活躍するようになっている [2] [3] [4] ．その例として，自動掃除機ロボット (図 2) が挙げられる．このロボットは，床の清掃作業にタスクとする．このロボットは，床の汚れを検出するセンサの他，物体との接触を検出するセンサや，転落を防止するために段差を検出するセンサが搭載されている．

図 2: 自動掃除機ロボット

さらに，多彩なセンサの開発により，1 台のロボットで複数のタスクを実行できるようになった．このことから，現在ではマルチタスク・ロボットの出現している．例えば，米国製の軍用遠隔操作多目的ロボットが挙げられる (図 3(a))．このロボットは，人間が立ち入るには困難な環境における複数の作業をタスクとする．実際に，このロボットは，原子力発電所における放射線量測定をタスクとして活動した．また，東京消防庁のレスキュー・ロボッ

トも挙げられる (図 3(b))．このロボットは，消防活動や救出活動をタスクとする．このロボットは，熱源を感知するために，温度を計測するセンサが搭載している．また，救助者を発見するためのカメラを搭載している．



(a) 遠隔操作多目的ロボット

(b) レスキュー・ロボット

図 3: マルチタスク・ロボット

一般に，ロボットがあるタスクを実行するためには，幾つかのセンサを搭載しなければならない．ロボットは，実行するタスクと自身が直面する環境に応じて，適応的な動作ができなければならないからである．特に，マルチタスク・ロボットは幾つかのタスクを実行するものである．そのため，マルチタスク・ロボットは，これまでのロボットよりも多くのセンサを搭載している．

1 体のロボットだけで複数のタスクを実行できることには，望ましいことが幾つかある．その 1 つは，ロボットの維持管理である．シングルタスク・ロボット—1 台で単一のタスクしか実行できないロボット— であれば，ロボットに実行させたいタスクが増えるたびに，その台数も増やさなければならない．このとき，それらを所持する人間は，その分だけ維持管理の手間を要する．一方，マルチタスク・ロボットであれば，その 1 台の維持管理さえすればよいため，その分の負担は軽減される．その他にも，マルチタスク・ロボットであれば，ロボットが占める空間を抑えることができる．シングルタスク・ロボットが幾つもあれば，その台数の分だけロボットは空間を占有してしまう．一方，マルチタスク・ロボットであれば，1 台分の空間さえ有ればよいため，ロボットが占める空間を抑えることができる．

したがって，今後はシングルタスク・ロボットのみならず，多くのセンサを搭載したマルチタスク・ロボットが幾つも出現することが予見される．

1.4 人間の予測によるセンサの選定とその難しさ

これまでは，ロボットに搭載するセンサは，人間による予測によって選定されていた．これは，ロボットが実行するタスクが簡単であるためや，ロボットが直面する環境が大きく変化しないために可能なことであった．そのため，

人間がロボットに搭載するセンサを事前に選定し，なおかつロボットがそれらを常に使い続けても，特に問題になることはなかったのである．

しかしながら，本研究では，ロボットがタスクを実行する上で必要となるセンサを特定することは難しいと考える．なぜなら，近年ではロボットが直面する環境が，より一層の複雑さを増しているからである．その複雑さの 1 つとして，ロボットが直面する環境の変化が考えられる．ロボットが直面する環境の変化によって，タスクを実行する上で必要となるセンサも異なることが考えられる．また，その他にも，ロボットが直面する環境が，人間にとっても未知・予測困難なものであることが考えられる．そのような環境へロボットを投入するとき，人間はそもそも必要となるセンサを特定すらできないことが考えられる．

前者の場合は，昼夜の明るさがその 1 つであると考えられる[4]．例として，ロボットが物体認識のためのカメラと超音波センサを搭載しているとする．昼間であれば，十分な太陽光が射し込むため，超音波センサよりもカメラによって物体を認識するのがよいと考えられる．一方，夜間であれば，僅かな月明かりしか射し込まないため，カメラによる物体の認識では問題が起こる可能性がある．よって，このときはカメラよりも超音波センサによって物体を認識するのがよいと考えられる．

後者の場合は，特に宇宙開発の現場が最たるものであると考えられる．例として，小惑星探査機のはやぶさが挙げられる [8] ．はやぶさは，自律航法誘導制御によって，小惑星イトカワに接近して着陸した．このとき，小惑星イトカワの正確な形や大きさ，位置は分かっていなかった．そのため，はやぶさに様々な状況の判断をさせる必要があったために，多くのセンサが搭載されていた．しかし，はやぶさが小惑星イトカワにタッチダウンするときに，本来であれば幾つかのセンサ情報を用いる予定であったものの，その内の加速度センサは用いなかったようである．その他にも，月・惑星探査ローバなどが挙げられる [9] [10].

1.5 ロボットの学習とセンサ情報の問題

前節の問題を解決するために，ロボットに可能な限り多くのセンサを搭載し，全てのセンサ情報をロボットのシステムに入力してしまうことと考えられる．しかしながら，そのような解決策が望ましくないことを本節で述べる．

近年では，ロボットに学習を適用する事例が数多くある [11] [12] [13]．これは，ロボットが未知の環境に直面しても，自律的に行動することでタスクを達成させるためである．

[4]このような例は，センサの測定方法であるパッシブ法とアクティブ法に依ることが，問題の 1 つであることが考えられる．パッシブ法は，測定対象から情報を得るために，その対象から発せられる物理量を受動的にセンシングする．一方，アクティブ法は，測定対象から情報を得るために，能動的にその対象へ物理量を発し，そこから戻ってくる物理量をセンシングする．この例では，カメラがパッシブ・センサ，超音波センサがアクティブ・センサである．

一般に，学習ロボットにおけるシステムの入力系も，その次数が増えるほど正確かつ適応的な出力ができる．すなわち，センサの種類や個数が多ければ多いほど，学習ロボットはより適応的に行動することができる．近年であれば，マルチタスク・ロボットのように幾つかのタスクを実行するために，その分だけ多くのセンサを搭載していることがある．このようなことから，学習を適用したマルチタスク・ロボットであれば，その全てのセンサ情報をロボットのシステムに入力してしまうことが考えられる．

しかしながら，いずれの学習の方法であっても，その入力系の次数が増えると何らかの問題が起こることが知られている．それは，多くの場合，その学習に要するメモリや計算時間といったコストの増大である．このような問題は,「次元の呪い」と呼ばれている（図 4）.

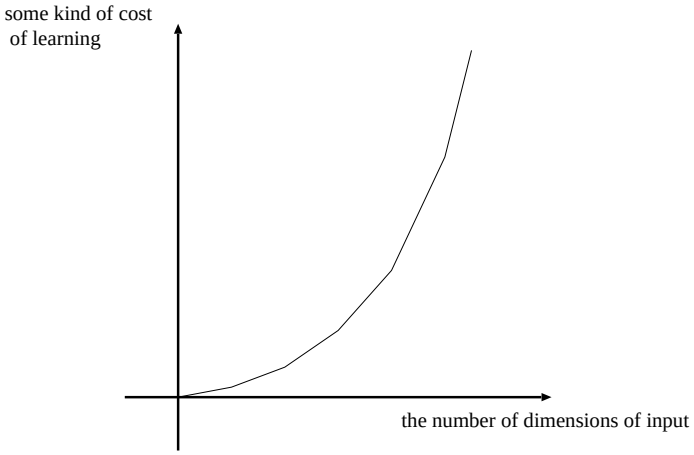


図 4: 学習制御における次元の呪い

ニューラル・ネットワークで起こる問題を例に挙げる．多くの研究では，3階層のネットワーク構造で構成されるニューラル・ネットワークの入出力の関係を，非線形関数として近似することがしばしば試みられている．この近似によって得られる解の精度を上げるためには，多くの基底関数を必要とする．そして，この基底関数に関するパラメータの個数は，入力系の次数に依存する．すなわち，入力系の次数—入力に相当するニューロンの個数—に比例して，多くのパラメータが必要になることが知られている．同様に，強化学習で起こる問題も例に挙げる．強化学習では，十分な回数だけ行動価値を更新しなければその値が定まらない．すなわち，学習が収束しない．この行動価値の個数は，状態と行動の組み合わせ—状態行動対—の数だけある．すなわち，入力系の次数—遷移可能な状態数—に比例して，学習に多くの時間を必要とすることが知られている．このニューラル・ネットワークのパラメータの個数と強化学習の学習時間は，いずれも入力系の次元数に応じて指数関数的に増大する．

したがって，このような問題を解決するために，ロボットに全てのセンサ情報を入力させることは，決して望ましくないことであると考えられる．特に，より多くのセンサを搭載するマルチタスク・ロボットに学習が適用されるとき，この問題は特に顕著に現れてしまうことが考えられる．

1.6 本研究の目的

本研究では，学習を適用したロボットが自律的にセンサ情報を選択し，かつ効率的に行動選択する方法を実現する．

ここで，近年では，ロボットに強化学習を適用する事例が多いことに着目する [14] [15] [16] [17]．強化学習は，未知の環境でも自律的に適応的な行動を獲得できる特長を持つ学習である．そこで，本研究では強化学習の枠組みの中で，ロボットが自律的にセンサ情報を選択する方法，選択したセンサ情報を利用する方法を考える．強化学習については，第 3 章で詳しく述べる．

1.7 従来研究

従来研究には，光永法明，浅田稔らの「移動体の意思決定のための情報量基準に基づく観測対象選択戦略」がある [18]．この研究では，行動決定木を情報量を基に作成することによりトレーニングデータの圧縮と観測順序の記述が統一的に扱えることを示している．

しかし，この研究による手法では，ロボットが周囲の場所を見渡せば，そのときに十分な情報が得られるという仮定に基づいているが，必ずしもそれが成立しない場合も考えられる．また，手法が何よりもランドマークに強く依存していることが考えられる．

1.8 アプローチの概要

本研究では，ロボットが自律的にセンサ情報を選択する方法を考える．この方法を実現するためには，まず，センサ情報の必要性を定量的・定性的に評価する指標を考えなければならない．そして，その指標に基づいてセンサ情報を必要・不要を判別しなければならない．特に，ロボットが環境の変化にも自律的に追従してセンサ情報を選択するために，環境を要因とする指標を考える．そこで，本研究では，環境の物理量とタスクの進捗度の間に相関が存在することに着目する（図 5）．

第 2 章では，この事実に基づいたアプローチについて詳しく述べる．

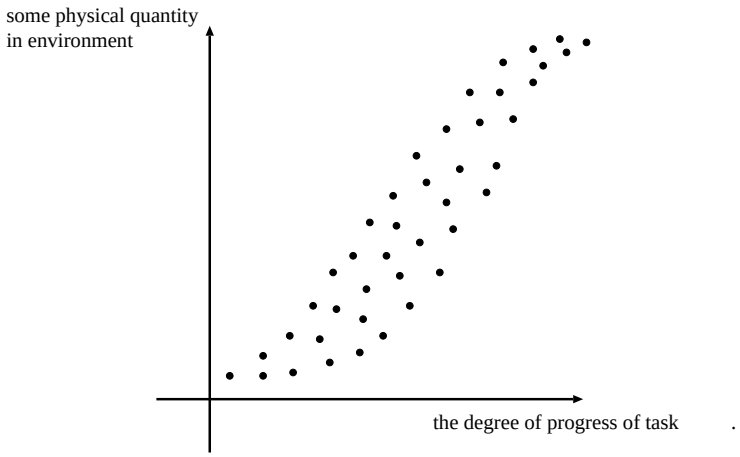


図 5: 環境の物理量とタスクの進捗度の相関

1.9 本論文の構成

本章では最後に，本論文の構成について述べる．第 1 章は，環境が変化することや環境が未知であることから，本研究ではロボットに搭載すべきセンサを特定する難しさに着目することを述べた．加えて，その問題解決が難しいことであることについて述べた．そして，本研究では，ロボットがセンサ情報を自律的に選択し，それにより効率的な行動選択する方法を強化学習の枠組みの中で実現することを述べた．

第 2 章では，本研究のアプローチについて詳しく述べる．ここでは，環境の物理量とタスクの進捗度の間に相関が存在することに着目する．

第 3 章では，本研究がロボットに適用する学習として，強化学習について述べる．ここでは，強化学習の枠組み中でセンサ情報を利用するとき，センサ情報は，強化学習における状態の定義に用いられるものであることを述べる．

第 4 章では，強化学習の枠組みの中で，ロボットがセンサ情報を自律的に選択する方法，および，その方法によって選択したセンサ情報により効率的に行動選択する方法を考える．そして，強化学習を基本とし，これらの方法と連係する学習機構について説明する．

第 5 章では，本研究が提案する方法の有効性を検証する実験について述べる．この章では，仮想環境における 2 つの実験と，実環境における実験の成果を示す．そして，これらの実験の結果から，一般的な強化学習と提案手法の性能について比較し，考察する．

第 6 章では，本論文の総括について述べる．ここでは，本研究における今後の課題についても述べる．

最後に，付録として，第 5 章で述べた実環境における実験で使用したロボットの図面を掲載する．ここでは，本実験のために我々が製作したロボットの

投影図と組立図について説明する．これらの付録は，実環境における実験の設定の説明を補完するためのものである．

以上の構成によって，本研究の有用性とその成果を示す．

2　タスクとセンサの関係

本研究では，第 4 章にてセンサ情報を自律的に選択する方法を提案する．そのための準備として，本章の第 1 節では，センサ情報の必要性を示す定量的な指標を考える．第 2 節では，センサ情報とタスクの達成度の相関について述べる．

2.1　タスクの進捗度と環境の物理量の関係

ロボットが自律的にセンサ情報を選択する方法を実現するためには，まず，センサ情報の必要性を定量的・定性的に評価する指標を考えなければならない．特に，ロボットが環境の変化にも自律的に追従してセンサ情報を選択するために，環境を要因とする指標を考える．

そこで，本研究では，環境の物理量とタスクの進捗度の間に相関が存在することに着目する．環境の物理量とタスクの進捗度の間には，相関があることが十分に考えられる．なぜなら，ロボットが実行するタスクの達成目標は，環境の物理量に依存して決定されるからである．そして，ロボットがタスクを実行するときには，ロボットが直面する環境の探索が伴うからである．このとき，ロボットは自身が直面する環境が，タスクの達成目標に相当する環境となるよう，アプローチ的な行動をする．このアプローチの過程で，環境の物理量は拡散しているため，連続的に変化することが考えられる．

この例として，火山調査をタスクとしてロボットに与えることを考える．すなわち，タスクを実行する環境は火山である．また，タスクの目的は，火山から発生するガスの化学組成の検出とする[5]．さて，このようなタスクを考えるにあたり，まずは，タスクの手順を以下の通りに分割する．

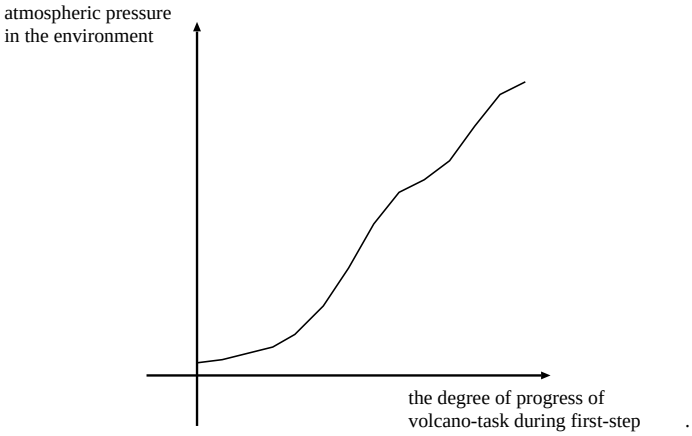
1. 火山の麓から，その頂上周辺への探索
2. 火山の頂上周辺から，火口付近への接近
3. 火山ガスの化学組成の検出

はじめに，ロボットが第 1 手順を実行するときを考える．このとき，ロボットが火山の麓から出発し，その頂上へ近付くにつれて，ロボットの周囲の大気圧は徐々に下がっていくだろう．なぜなら，大気圧は，山の高度に比例して拡散しているからである．すなわち，このタスクにおける第 1 手順では，その進捗度が大きくなるにつれて，ロボットの周囲の大気圧が下がることが考えられる．次に，ロボットが第 2 手順を実行するときを考える．先の手順により，ロボットは火山の頂上周辺に到着しており，そこでは既に火山ガスが

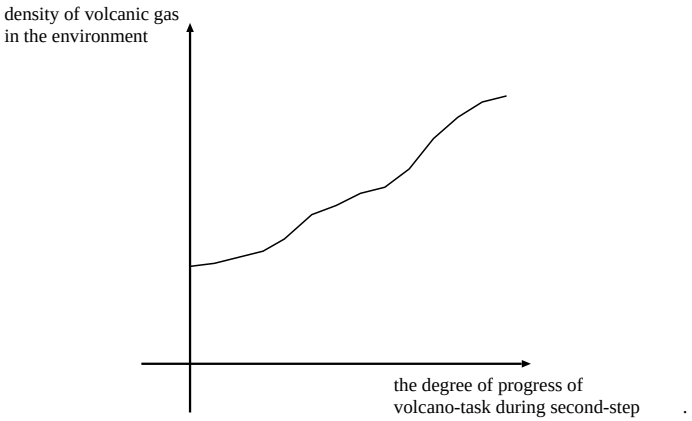
[5]火山が噴火するときの前兆現象の 1 つに，火山ガスの化学組成の変化がある．火山ガスは，多くの場合，人体に有毒である硫化水素や二酸化硫黄，塩化水素を含む．このような火山ガスを吸入して，その場で人間が死亡してしまうことも多々ある．したがって，このような危険を伴う調査活動は，人間に代わってロボットが担うべきであろう．

ある程度だけ検出されていることが考えられる．このとき，ロボットが頂上周辺から火口付近へ近付くにつれて，火山ガスの濃度は徐々に高くなっていくだろう．なぜなら，火山ガスは，火口からの距離に比例して拡散しているからである．すなわち，このタスクにおける第 2 手順では，その進捗度が大きくなるにつれて，ロボットの周囲の火山ガスの濃度が高くなることが考えられる．これらの手順を踏まえて，ロボットはタスクの目的である第 3 手順を実行できる．以上より，火山調査をタスクとするこの例では，その進捗度と大気圧および火山ガスの濃度のそれぞれに正の相関があると言える (図 6)．

したがって，ロボットはタスクを実行するときに自身が直面する環境を探索することと，環境の物理量が拡散していることにより，環境の物理量とタスクの進捗度の間には相関が存在すると考えられる．



(a) 第 1 手順の進捗度と大気圧の大きさ



(b) 第 2 手順の進捗度と火山ガスの濃度

図 6: 火山調査タスクの進捗度と大気圧および火山ガスの濃度の関係

2.2　タスクの進捗度とセンサ情報の相関

タスクの進捗度とセンサ情報の間にも何らかの相関があると考えられる．なぜなら，ロボットは環境の物理量をセンサを通して入力するからである．すなわち，ロボットは環境の物理量そのものを実際に知覚しているのではなく，それに対応するセンサ情報によって環境を認識している．よって，タスクの進捗度とセンサ情報との間の相関は，タスクの進捗度と環境の物理量との間の相関に対応すると考えられる．

しかしながら，タスクの進捗度と環境の物理量の相関と，タスクの進捗度とセンサ情報の相関は，必ずしも等価ではない．これらの関係が等価であるには，環境の物理量とセンサ情報の関係式が 1 次式 $y = x$ の関係であることが条件である．したがって，これらの関係式の傾きの正負が逆であれば，それらの 2 つの相関の正負も逆となることが考えられる．また，これらの関係式が非線形であれば，それらの 2 つの相関は大きく異なることが考えられる．これは，本論文の第 5 章で述べる赤外線センサ GP2Y0A21YK0F がその例である[6]．

環境の物理量とセンサ情報の関係式は，千差万別である．すなわち，環境の物理量からセンサ情報への変換は一様ではない．なぜなら，センサはそれぞれ固有の計測原理に従って環境の物理量を測定し，固有の電気信号を出力するからである．また，ロボットの A/D コンバータが量子化するときの量子数も，その性能によって異なるからである．

よって，ロボットがセンサ情報を自律的に選択するとき，タスクの進捗度とセンサ情報の相関を求めるのは適切ではない恐れがあることに留意しなければならない．タスクの進捗度と環境の物理量の間の相関を利用するには，センサと A/D コンバータの固有の変換関係に従って，センサ情報から環境の物理量を逆変換して求めなければならない．

しかしながら，近年では，画像認識や音声認識に使用されるセンサのように，1 つのセンサが処理する情報量が極めて大きいことがある．そのため，センサ情報から環境の物理量への逆変換には，多大な計算時間が求められることが予測される．

したがって，本研究では，タスクの進捗度と環境の物理量の間の相関とタスクの進捗度とセンサ情報の間の相関が等価ではなくても，後者の相関の強さが大きく失われることがなければ，後者の相関を調べることを許容するものとする．本研究では，これらのことを踏まえた上で，タスクの進捗度とセンサ情報の相関に基づいた手法を第 4 章で提案する．

[6]赤外線センサ GP2Y0A21YK0F の電圧-距離の特性関係については，図 25 を参照されたい．

3　強化学習とセンサの関係

本章では，本研究の提案手法を説明する準備として，強化学習について述べる．はじめに，第 1 節では，強化学習の概要について述べる．ここで，強化学習が人間や動物の学習の過程をモデルとしていることや，そのことによる強化学習の特長について述べる．次に，第 2 節では，強化学習におけるエージェントの行動選択と行動評価について簡単に述べる．第 3 節では，強化学習における状態について述べる．最後に，第 4 節でセンサ情報による状態の定義について述べる．特に，第 4 節で述べるセンサ情報の定義は，ロボットの効率的な行動選択を実現する方法で重要となる．

3.1　強化学習の概要

強化学習とは，学習者が環境との相互作用を通じて目的の行動を獲得する学習である.「スキナー・ボックス」や「パブロフの犬」などが挙げられるように，強化学習は人や動物が学習する過程をモデルとしている．

スキナー・ボックスを例として述べる．この例ではネズミが学習者である．ある箱の中に，1 匹のネズミが入れられる．この箱の中には 1 本のレバーが取り付けられている．このレバーを引くと，そのネズミの好物の餌が出される．ネズミがその箱の中で動き回っていると，あるときにネズミはこのレバーを偶発的に引く．これにより，ネズミは好物の餌を獲得することができる．さらに同じことを経験すると,「レバーを引く」という行動が強化される．これを繰り返すことで，ネズミは「レバーを引けば，好物の餌が出る」という経験的な知識を学習する．ネズミはこの知識を利用することで，さらに好物の餌を獲得することができる．こうして，ネズミはその箱の中で「レバーを引く」という行動を獲得できるのである．

強化学習は，このような学習の過程をモデルとする．ここで，強化学習では，学習者に行動を教示する教師は存在しないという特長がある．これによって，エージェントは自身が存在する環境が未知である場合でも，自律的に目的を達成できる．強化学習の過程をモデル化した図 7 に示す.

強化学習における学習者は，エージェント（Agent）と呼ばれる．このエージェントが相互作用を行う対象は，環境（Environment）と呼ばれる．両者は学習の過程で継続して相互作用する．このとき，エージェントは自発的に行動（Action）することで環境に作用する．環境は，その行動に応じて，状態（State）と報酬（Reward）を応答として返す．状態とは，エージェントが直面する環境の状況を表現するものである．また，報酬とは，その状態や行動の望ましさを表現するスカラ値である．先述したスキナー・ボックスの例で言えば，ネズミの好物の餌に相当する．エージェントは，このような環境との相互作用を試行錯誤に繰り返すことで，より多くの報酬を獲得することを目的とする．

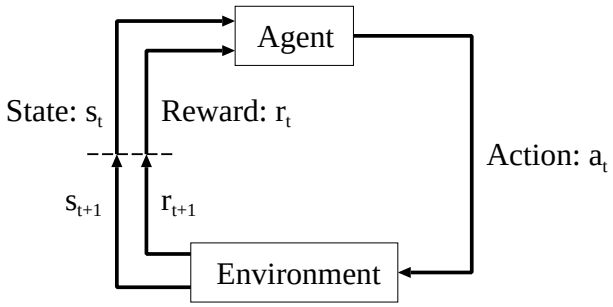


図 7: 強化学習の過程

強化学習では，これらの行動，状態，報酬に加えて，価値という要素によって学習の振る舞いを実現する．先のスキナー・ボックスのネズミのように，ある行動に対する報酬が大きければ，エージェントはその行動をさらに繰り返すようになるべきである．言い換えれば，そのときの行動が強化されるべきである．逆に，ある行動に対する報酬が小さければ，エージェントはその行動を控えるようになるべきである．言い換えれば，そのときの行動が弱化されるべきである．価値は，こうした強化や弱化を実現するために，報酬が行動や状態の即時的な望ましさを表現しているのに対し，状態や行動の長期的な望ましさを表現している．すなわち，その後に続きそうな状態群とそれらの状態群で得られそうな報酬を考慮した望ましさである．エージェントは環境から報酬が与えられる度に，この価値を評価していく．そして，エージェントはこの価値に従って行動を選択することで，結果として最大の報酬を得ることができるのである．これにより，エージェントは目的の行動を獲得することに繋がるのである．

3.2 行動選択と行動評価

前節では，強化学習の概要について述べた．以上を踏まえて，本節では，強化学習における行動選択と行動評価について述べる．さらに，本節ではそれぞれの代表的な方法も挙げて説明する．ここで述べた方法は，第 5 章の実験でエージェントに適用するものである．

はじめに，強化学習におけるエージェントの行動選択について述べる．前節では，エージェントは自発的に行動することを述べた．このとき，エージェントは過去に試した行動よりも更に良い行動を発見するために，確率的に行動することがある．強化学習では，このことを探査と呼ぶ．逆に，過去に試した行動の中で最善な行動を選択することを，知識利用と呼ぶ．最善な行動とは，その状態における選択可能な行動の中で，最も高い価値を有するものである．また，知識とは，エージェントが学習の過程で獲得した状態や行動

の価値のことを指す．エージェントは，この探査と知識利用を確率的に使い分けて行動を選択する．

これを実現する簡単な方法に，$\epsilon-$greedy 法がある．この方法は，確率 ϵ で探査し，確率 $1-\epsilon$ で知識を利用するものである．ここで，現在の状態を s，その状態 s で選択可能な行動の集合，および任意の行動をそれぞれ $A(s), a$，その状態 s においてある行動 a 選択する価値を $Q(s,a)$ とおく．このとき，この方法による行動選択 π は，次式の通りに書ける．

$$\pi(s) = \begin{cases} \arg\max_{a \in A(s)} Q(s,a) & (\text{確率 } 1-\epsilon) \\ \forall a \in A(s) & (\text{確率} \epsilon) \end{cases} \tag{1}$$

このように，行動選択 π は，遷移可能な状態群のある状態 $s \in S$ から，その状態における選択可能な行動群の行動 $a \in A(s)$ への確率的な写像で実現される．

次に，強化学習におけるエージェントの行動評価について述べる．前節では，エージェントは結果として最大の報酬を得るために，価値に従って行動することを述べた．エージェントはこの価値を推定するために，報酬や価値などの値をもとにして，漸進的に計算する．エージェントはこの計算を何ステップか繰り返すことによって，より確からしい価値—真の価値—を求める．

これを実現する代表的な方法に，1 ステップ Q 学習がある．この方法は，遷移後の状態における報酬とその状態における最大の価値をもとに，最適な行動価値を近似するものである．ここで，現時刻 t の状態と次時刻 $t+1$ —遷移後— の状態をそれぞれ s_t, s_{t+1}，ある状態 s における行動 a の価値を $Q(s,a)$ とおく．このとき，この方法による行動評価は，次式の通りに書ける．

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_{a \in A(s_t)} Q(s_{t+1}, a) - Q(s_t, a_t) \right] \tag{2}$$

ここで，α はステップサイズ・パラメータ，γ は割引率と呼ばれる定数であり，それぞれ区間 $(0,1]$ の値をとる．式を見てわかる通り，α は 1 ステップ当たりの価値の更新量を調整するものである．また，γ は遷移後の状態における価値のフィードバックを調整するものである．このように，行動評価は，遷移可能な状態群のある状態 $s \in S$ において選択可能な行動群の任意の行動 $a \in A(s)$ が有する価値 $Q(s_t, a_t)$ の漸近式で実現される．

余談ではあるが，行動評価の式は漸近式ではなく，もともとは一般式の形で記述されていたものである．例えば，報酬の標本平均によって行動価値を推定する式は，一般式で

$$Q(a) = \frac{r_1 + r_2 + \cdots + r_{k_a}}{k_a}$$

と書ける．ここで，k_a は行動 a を選択した回数，$r_1, r_2, \cdots, r_{k_a}$ はその回数のときに獲得した報酬である．このような一般式だと，これらの値を全て格

納するためのメモリが必要となる．さらに，選択した回数 k_a が増えるたびにメモリも多く必要となる．しかしながら，これを漸化式にすれば

$$Q(a) \leftarrow Q(a) + \frac{1}{k_a + 1}\left[r_{k+1} - Q(a)\right]$$

と書ける．すなわち，行動 a の価値 $Q(a)$ とその更新回数 k_a，新たに獲得した報酬 r_{k+1} さえ分かればよい．このことは，有限な情報処理能力—演算速度，記憶容量—しか持たないロボットにとって重要なことである．さらに言えば，ロボットはしばしばマイコンのような組込み機器で構成される．すなわち，これらの機器の情報処理能力は，一般に我々が使用する PC よりも劣るものである．したがって，行動価値の更新式は漸化式の形で書かれることは，重要なことであると言える．

3.3 強化学習における状態

強化学習におけるエージェントは，環境との相互作用で環境の状態を認識する．この状態は，エージェントが直面する状況を決定するものである．強化学習では，状態の定義やその性質を問題とすることがある．本節では，この状態について説明する．

強化学習における状態の定義の方法は，特段に制限を受けるものではないとされている [19]．すなわち，センサ情報の入力のような低レベルで具体的な知覚でもよければ，その環境に存在する物体を記号によって記述するような高レベルで抽象的な知覚でもよい．さらには,「驚いている」というような観念的なものでも定義してよいとされている．このように，強化学習における状態の定義はとても自由なものである．

しかしながら，強化学習における状態の定義では，マルコフ性と呼ばれる環境の性質に注意を払うことが多い．マルコフ性とは，過去の状態に全く依らず，現在の状態のみにしか依存しないという性質である．強化学習では，エージェントが意思決定—行動を選択—するとき，現在の状態にしか基づかないものとしている．これにより，エージェントは現在の状態のみを根拠として，将来の状態と期待される報酬の全てを予測することができる．これは，強化学習におけるエージェントが自身の経験のみによって自律的に行動するための重要な性質である．マルコフ性を満たす環境は，マルコフ決定過程（MDP：Markov decision process）と呼ばれる．状態と行動の空間が有限である場合，有限マルコフ決定過程（有限 MDP）と呼ばれる．

以上のことを踏まえて，後節では，強化学習における状態をセンサ情報によって定義する．

3.4 センサ情報による状態の定義

本節ではセンサ情報による状態の定義について述べる．ここでは，エージェントが n 個のセンサを搭載していることを前提とし，これらのセンサ情報によって状態を定義することを考える．

はじめに，n 個のセンサ情報 $e_1, e_2, \ldots, e_n$ による単純な状態の定義を考える．ここで，任意の時刻 t における n 個のセンサ情報を $e_{1,t}, e_{2,t}, \ldots, e_{n,t}$ とおく．また，$e_{1,t}, e_{2,t}, \ldots, e_{n,t}$ がそれぞれ取り得る値から成る集合を $E_1, E_2, \ldots, E_n$ とおく．このとき，時刻 t におけるエージェントの状態 s_t は，次式の通りに定義することが考えられる．

$$s_t := \{(e_{1,t}, e_{2,t}, \ldots, e_{n,t}) | e_{1,t} \in E_1, e_{2,t} \in E_2 \ldots, e_{n,t} \in E_n\} \tag{3}$$

これは，センサ情報の列ベクトルで定義されたものであり，ごく簡単な定義であるといえる．

しかしながら，このような単純な定義を適用することはあまり適当ではないと考えられる．なぜなら，センサが出力する電気信号は，A/D コンバータによって大きい数で量子化されるからである．すなわち，それぞれのセンサ情報の集合の濃度 $|E_1|, |E_2|, \ldots, |E_n|$ がとても大きいからである．例えば，あるA/D コンバータは，0(V) から 5(V) の範囲で，センサが出力する電気信号を 1024 の数だけ量子化する．エージェントが搭載する n 個のセンサが，いずれも 0(V) から 5(V) の全ての範囲で何らかの信号を出力するならば，すなわち，$|E_1|, |E_2|, \ldots, |E_n|$ の濃度がいずれも 1024 であるならば，その状態数は 1024^n となる．これでは，エージェントは現実的な時間内で学習することはできないことが考えられる．

そこで，本研究では，センサ情報をそのセンサが計測する物理量に変換し，その物理量の任意長の固定区間から状態変数に写像することを考える．ここで，センサ情報の任意長の固定区間から写像しないのは，センサ情報が物理量に対して線形であるとは限らないからであることに注意されたい．

はじめに，あるセンサが計測する任意の物理量を p，物理量 p が取り得る値の集合を P とおく．このとき，あるセンサ情報 $e \in E$ から物理量 $p \in P$ に変換する写像 f を次式の通りに定義する．

$$f : E \mapsto P \tag{4}$$

この具体的な式は，センサの測定原理とその特性によって異なるものである．したがって，ここでは単に写像の関係だけを定義するものに留める．

次に，写像 f により求めた任意の物理量 $p \in P$ の任意長 L の固定区間から，ある状態変数 u に変換する写像 g を，次式の通りに定義する．

$$\begin{cases} g : P \mapsto U \\ g(p) = \{u = i | (i-1) \times L \leq p < i \times L, i \in N\} \end{cases} \tag{5}$$

ここで，N は自然数の集合である．これにより，任意の物理量 $p \in P$ を更に自然数 $u \in N$ で離散化することができる．

これより，任意のセンサ情報 $e \in E$ から合成写像 $f \circ g$ によって変換される状態変数 u によって状態を定義する．ここで，任意の時刻 t における n 個のセンサ情報を $e_{1,t}, e_{2,t}, \ldots, e_{n,t}$ ，それらに対応する状態変数が $u_{1,t}, u_{2,t}, \ldots, u_{n,t}$ とおく．このとき，エージェントの状態 s_t を次式の通りに定義する．

$$s_t := \{(u_{1,t}, u_{2,t}, \ldots, u_{n,t}) | u_{i,t} = f \circ g(e_{i,t}), i \in N\} \tag{6}$$

これにより，写像 g における物理量 P の固定長 L を調整することで，等間隔かつ適切な数だけ，エージェントは状態を認識することができると考えられる．

4　提案手法

本章では，第 1 節で，本研究が提案する学習機構の設計概念を示す．この学習機構は強化学習を基本とすることを述べる．第 2 節では，ロボットが自律的にセンサ情報を選択する方法について述べる．ここでは，その方法として，強化学習における報酬とセンサ情報を 2 変量とする相関分析によって実現することを述べる．これにより，ロボットは選択したセンサ情報による状態認識を実現することができる．第 3 節では，その状態認識によってロボットが効率的に行動選択する方法について述べる．ここでは，先述した方法による状態認識に応じて行動価値を利用することを述べる．これにより，ロボットは効率的な行動選択を実現することを述べる．最後に，第 4 節では，これらの方法と強化学習を組み合わせた学習機構の具体的な流れについて述べる．ここでは，これらの方法が強化学習の過程で，いつ，どのように機能するかを述べる．

4.1　提案する学習機構の概念

本研究では，強化学習の枠組みの中で，ロボットが自律的にセンサ情報を選択する方法，および，ロボットが効率的に行動を選択する方法を考える．はじめに，本研究が提案する学習機構の設計概念を図 8 に示す．

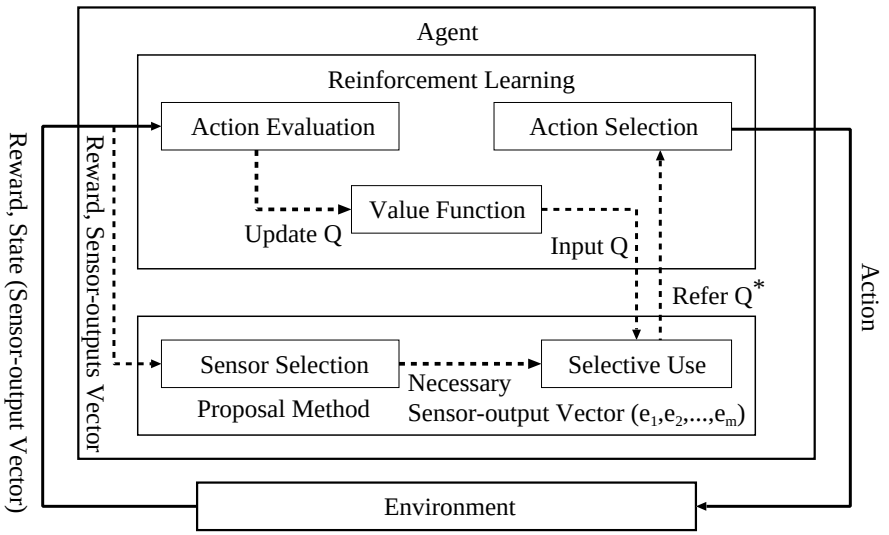


図 8: 本研究が提案する学習機構の設計概念

図 8 の通り，本研究が提案する学習機構は，強化学習を基本とするものである．そして，本研究が提案する 2 つの手法が，強化学習と付加的に連係し

て機能する．これらの提案手法は，強化学習における幾つかのパラメータを必要とするものである．

Sensor Selection というモジュールは，ロボットが自律的にセンサ情報を選択することを実現する機能である．これは，センサ情報による選択的な状態認識を実現するものである．Selective Use というモジュールは，ロボットが効率的に行動を選択することを実現する機能である．これは，その選択的な状態認識に基づいた行動選択を実現するものである．

これらの 2 つの方法と強化学習を連係させることで，ロボットがより効率的にタスクを達成できる学習機構の確立を目指す．本節に続く 2 つの後節では，これらの具体的な方法について述べる．そして，本章の最後の節では，本研究が提案する学習機構の具体的な流れについて述べる．

4.2 センサ情報の自律的な選択方法

本節では，ロボットが環境の変化に追従して自律的にセンサ情報を選択する方法を考える．この方法により，ロボットが自律的に選択したセンサ情報による状態認識を実現する．

本研究では，タスクの進捗度と相関が見られるセンサ情報は，ロボットがタスクを実行する上で必要であると考える．そこで，まずはセンサ情報の必要性を定量的に評価する方法を考える．第 2 節で述べた通り，本研究ではその方法として，センサ情報とタスクの進捗度を相関分析することを考える．すなわち，センサ情報の必要性を示す値は，相関係数に相当する．

ここで，タスクの進捗度は，強化学習における即時的な報酬値によって表現するものとする．第 3 章で述べた通り，報酬は，その状態や行動の望ましさを表現するスカラ値である．この報酬の与え方には，即時的なものと遅延的なものの 2 通りがある．即時的な与え方は，いずれの状態においても例外なく報酬を与えるものである．遅延的な与え方は，目標とする状態やその経過にあたる状態においてのみ報酬を与えるものである．いずれの与え方であっても，目標とする状態では最大の報酬が与えられる．特に，即時的な与え方について言えば，目標の状態の近さに応じて高く報酬を与えることがある．このことから，報酬の値とタスクの進捗度の値の傾向と性質は近しいと言える．これより，本研究では，タスクの進捗度は報酬値によって表現できると考える．

これより，センサ情報と報酬値を 2 変量とする相関分析の式を考える．前提として，ロボットは n 個のセンサを搭載しているとする．ここで，任意の時刻 t における n 個のセンサ情報を $e_{1,t}, e_{2,t}, \ldots, e_{n,t}$ とおく．また，$e_{1,t}, e_{2,t}, \ldots, e_{n,t}$ が現時刻 $T(T \geq 0)$ までに $e_{1,t}, e_{2,t}, \ldots, e_{n,t}$ が取得した値から成る集合を $E_{1,T}, E_{2,T}, \ldots, E_{n,T}$ とおく．また，任意の時刻 t における即時報酬を r_t ，現時刻 T までにエージェントが獲得した報酬値の集合を R_T とおく．このと

き，$E_{i,T}(1 \leq i \leq n)$ と R_T を 2 変量とする相関係数 $\rho_{i,T}$ は，次式に従って求めることができる[7]．

$$\rho_{i,T} = \frac{\sum_{e_i \in E_{i,T}} (e_{i,t} - \overline{e}_{i,T})(r_t - \overline{r}_{t,T})}{\sqrt{\sum_{e_i \in E_{i,T}} (e_{i,t} - \overline{e}_{i,T})^2} \sqrt{\sum_{r_t \in R_T} (r_t, -\overline{r}_T)^2}} \quad (7)$$

$$\left(\overline{e}_{i,T} = \sum_{e_{i,t} \in E_{i,T}} e_i \quad , \quad \overline{r}_T = \sum_{r_t \in R_T} r_t \right)$$

この式に従い，全てのセンサ情報 $E_{1,T}, E_{2,T}, \cdots, E_{n,T}$ について相関分析することで，個々のセンサ情報の必要性を定量的に評価することができると考えられる．

次に，センサ情報の必要性を示す値として得た相関係数 $\rho_{1,T}, \rho_{2,T}, \cdots, \rho_{n,T}$ によって，必要か否かを判別する方法を考える．本研究では，$\rho_{1,T}, \rho_{2,T}, \cdots, \rho_{n,T}$ の内で，その値が任意に設定した閾値 ρ 以上のものを必要とし，それ未満のものは不要とする．すなわち，$\rho_{i,T} \geq \rho$ であるとき，その時刻 T におけるセンサ情報 $e_{i,T}$ を必要とする．

この方法を実現するにあたり，ロボットは各時刻毎のセンサ情報とそのときに獲得した報酬値を記憶する必要がある．そのため，ロボットはこれらを記憶する知識テーブルを持つものとする．ここで，任意の時刻 t におけるセンサ情報 $e_{1,t}, e_{2,t}, \cdots, e_{n,t}$ および報酬値 r_t の組み合わせを，レコードと呼ぶことにする．このとき，現時刻 $T(T \geq 0)$ における知識テーブルは，表 1 の通りに定義する[8]．

本研究が提案する手法を適用したロボットは，時刻に昇順するようにして，上の知識テーブルに新たなレコードを追加する．ただし，知識テーブルに存在するレコードの内で，時刻を除くレコードの項—列についてのデータ—の組み合わせがいずれかのレコードと重複するとき，そのレコードは追加されないものとする．すなわち，現時刻 T までに追加されなかったレコード数が $m(m \leq T+1)$ であれば，その時刻の知識テーブルに存在するレコード数は $T+1-m$ である．本研究の提案手法を適用したロボットは，全ての時刻毎にこのような知識テーブルを参照して，それぞれのセンサ情報について相関係数を求めていくものとする．

---

[7]この式に従って求める相関係数 ρ は，ピアソンの積率相関係数と呼ばれる．この相関係数は，2 変量の関係がどれだけ線形関係に近しいかを示す値である．ただし，相関関係によって因果関係を説明することはできないことに注意されたい．すなわち，2 変量 A, B の間に相関があるとき，$a \in A$ が高くなれば $b \in B$ が高くなるというわけではない．2 変量 A, B は，一方を説明変数，他方を目的変数として区別して見ることができず，そのどちらにも考えることができるのである．

[8]本論文ではテーブルと書いたが，時刻 T までの全てのレコードが参照できれば，必ずしもそのデータ構造がその通りである必要はない．すなわち，2 次元方向のデータ構造である必要はない．実際，本研究ではこれを単方向のリストによって実現している．

表 1: 現時刻 T における自律的選択に関する知識テーブル

時刻 t	センサ情報 $E_{1,T}$	センサ情報 $E_{2,T}$	$\cdots$	センサ情報 $E_{n,T}$	報酬値 R_T
0	$e_{1,0}$	$e_{2,0}$	$\cdots$	$e_{n,0}$	r_0
1	$e_{1,1}$	$e_{2,1}$	$\cdots$	$e_{n,1}$	r_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
T-1	$e_{1,T-1}$	$e_{2,T-1}$	$\cdots$	$e_{n,T-1}$	r_{T-1}
T	$e_{1,T}$	$e_{2,T}$	$\cdots$	$e_{n,T}$	r_T

ここで，本研究では，この知識テーブルの追加についてこれ以上のルールを設けていない．すなわち，ある時刻におけるレコードは，知識テーブルに存在するものと重複しない限り，際限無く追加されるものとしている．これは，ロボットが持つメモリの容量が有限であることを考えると，望ましいことではない．本来であれば，任意の時刻におけるレコードを適切な基準によって選択し，それを削除するルールを設けるべきである．特に，その最大数はA/D コンバータが量子化する数に依るが，実環境においては多くの個数のセンサ情報が認識されると考えられる．すなわち，知識テーブルに記憶されるレコード数が甚大になることが考えられる．しかしながら，本研究では，これらのことについての議論はこれ以上しない．

4.3 センサ情報の選択的な利用方法

本節では，ロボットが前節で述べた方法によって必要と選択したセンサ情報を，さらに選択的に利用する方法を考える．この方法により，ロボットが自律的に選択したセンサ情報による効率的な行動選択を実現する．

さて，強化学習の枠組みでセンサ情報を利用する方法の 1 つとして，センサ情報によって状態を定義することを第 3 章で述べた．そこで，まずはロボットが必要と選択したセンサ情報によって，強化学習における状態を再定義することを考える．前提として，ロボットは n 個のセンサを搭載しているとする．ここで，任意の時刻 t における n 個のセンサ情報を $e_{1,t}, e_{2,t}, \ldots, e_{n,t}$ とおく．また， $e_{1,t}, e_{2,t}, \ldots, e_{n,t}$ が取り得る値の集合を $E_1, E_2, \ldots, E_n$ とおく．また，先述した方法によって，現時刻 T における m 個のセンサ情報 $e_{1,T}, e_{2,T}, \cdots, e_{m,T} (0 \leq m \leq n)$ が必要，残りの $n-m$ 個のセンサ情報 $e_{m+1,T}, e_{m+2,T}, \cdots, e_{n-m,T}$ が不要と選択されたとする．このとき，現時刻 T における状態 s_T の定義を，(3) 式に従うならば (8) 式の通りに，(6) 式に従うならば (9) 式の通りに状態 s_T^* として再定義するものとする．

$$s_T^* := \{(e_{1,T}, e_{2,T}, \ldots, e_{m,T}) | e_{1,T} \in E_1, e_{2,T} \in E_2 \ldots, e_{m,T} \in E_m\} \qquad (8)$$

$$s_T^* := \{(u_{1,T}, u_{2,T}, \ldots, u_{m,T}) | u_{i,T} \in N, 1 \leq i \leq m\} \qquad (9)$$

次に，状態 s_T における任意の行動 a の価値 $Q(s_T, a)$ を，再定義した状態 s_T^* に応じて低次元化することを考える．すなわち，s_T^* における任意の行動 a の価値 $Q^*(s_T, a)$ を計算することを考える．この低次元化は，状態 s_T における任意の行動 a の価値 $Q(s_T, a)$ の評価回数 $N(s_T, a)$ を重みとする加重平均式による計算で実現する．

現時刻 T における状態 s_T の定義が (3) 式に従うときを考える．ここで，現時刻 T においてロボットが必要と判別した m 個のセンサ情報の集合を $E_{1,T}, E_{2,T}, \cdots, E_{m,T} (0 \leq m \leq n)$ ，不要と判別した $n-m$ 個のセンサ情報の集合を $E_{m+1,T}, E_{m+2,T}, \cdots, E_{n,T}$ とおく．さらに，それらの不要なセンサ情報の集合の直積 $E_{m+1,T} \times E_{m+2,T} \times \cdots \times E_{n,T}$ を $\bar{E_T}$ ，その任意の列 $(e_{m+1}, e_{m+2}, \ldots, e_n)$ を $\bar{e}$ とおく．このとき，s_T^* における任意の行動 a の価値 $Q^*(s_T, a)$ を (10) 式の通りに計算するものとする．

$$Q^*(s_T^*, a) = \frac{\sum_{\bar{e} \in \bar{E_T}} N(s_T, a) Q(s_T, a)}{\sum_{\bar{e} \in \bar{E_T}} N(s_T, a)} \qquad (10)$$

$$\left(= \frac{\sum_{\bar{e} \in \bar{E_T}} N(s_T^*, \bar{e}, a) Q(s_T^*, \bar{e}, a)}{\sum_{\bar{e} \in \bar{E_T}} N(s_T^*, \bar{e}, a)} \right)$$

次に，現時刻 T における状態 s_T の定義が (6) 式に従うときを考える．ここで，現時刻 T においてロボットが不要と判別した $n-m$ 個のセンサ情報の集合を $E_{m+1,T}, E_{m+2,T}, \cdots, E_{n,T}$ ，それらのセンサ情報の集合に対応する状態変数の集合を $U_{m+1,T}, U_{m+2,T}, \cdots, U_{n,T}$ とおく．さらに，それらの状態変数の集合の直積 $U_{m+1,T} \times U_{m+2,T} \times \cdots \times U_{n,T}$ を $\bar{U_T}$ ，その任意の列 $(u_{m+1}, u_{m+2}, \ldots, u_n)$ を $\bar{u}$ とおく．このとき，s_T^* における任意の行動 a の価値 $Q^*(s_T, a)$ を (11) 式の通りに計算するものとする．

$$Q^*(s_T^*, a) = \frac{\sum_{\bar{u} \in \bar{U_T}} N(s_T, a) Q(s_T, a)}{\sum_{\bar{u} \in \bar{U_T}} N(s_T, a)} \qquad (11)$$

$$\left(= \frac{\sum_{\bar{u} \in \bar{U_T}} N(s_T^*, \bar{u}, a) Q(s_T^*, \bar{u}, a)}{\sum_{\bar{u} \in \bar{U_T}} N(s_T^*, \bar{u}, a)} \right)$$

ロボットは上式に従って，現時刻の状態 s_T において選択可能な任意の行動 $a \in A(s_T)$ について，状態 s_T^* における行動価値 $Q^*(s_T, a)$ を計算するものとする．そして，ロボットはそのときに得られた任意の行動 a の価値 $Q^*(s_T^*, a)$ を参照して，行動選択するものとする．ただし，全てのセンサ情報を不要と判別した場合，ロボットは本来の状態 s_T における任意の行動 a の価値 Q を参照して，行動選択するものとする．逆に，全てのセンサ情報を必要と判別した場合，$Q^*(s_T^*, a)$ は $Q(s_T, a)$ と等価になるため，同様にして任意の行動 a の価値 $Q(s_T, a)$ を参照して，行動選択するものとする．

これにより，強化学習の枠組みの中で，ロボットが自律的に選択したセンサ情報によって，さらに効率的に行動を選択することが実現できる．ただし，この方法によってロボットが行動選択しても，その後に評価される—更新される—行動価値は，本来の状態 s_T における価値 $Q(s_T, a_T)$ であるものとする．

この方法を実現するにあたり，ロボットは状態行動対毎の行動価値 Q の更新回数 N を記憶する必要がある．そのため，ロボットはこれらを記憶する知識テーブルを持つものとする．ここで，現時刻 T までに遷移した k_1 個の状態を $s_1, s_2, \ldots, s_{k_1}$，およびそれらの状態で選択可能な k_2 個の行動を $a_1, a_2, \ldots, a_{k_2}$ とおく．このとき，その知識テーブルは表 2 の通りに定義する．

表 2: 選択的利用に関する知識テーブル

状態 s	行動 a_1	行動 a_2	$\cdots$	行動 a_{k_1}
s_1	$N(s_1, a_1)$	$N(s_1, a_2)$	$\cdots$	$N(s_1, a_{k_2})$
s_2	$N(s_2, a_1)$	$N(s_2, a_2)$	$\cdots$	$N(s_2, a_{k_2})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
s_{k_1}	$N(s_{k_1}, a_1)$	$N(s_{k_1}, a_2)$	$\cdots$	$N(s_{k_1}, a_{k_1})$

この知識テーブルでは，経験した状態行動対であれば，どのレコードについても例外なく記憶しなければならない．したがって，ロボットに本研究が提案する手法を適用する場合，ロボットは，少なくとも上の知識テーブルを記憶するだけのメモリの容量を要求される．また，それと同時に，この知識テーブルに従って行動価値 Q^* を計算しなければならない．

4.4 提案する学習機構

本節では，第 2 節，第 3 節で述べた方法を踏まえて，本研究が提案する学習機構の具体的な流れについて述べる．ここでは，これらの方法が強化学習の過程で，いつ，どのように機能するかを述べる．

はじめに，本研究が提案する具体的な学習機構を表現した図を図 9 に示す．Correlation Analysis というモジュールは，Sensor Selection というモジュー

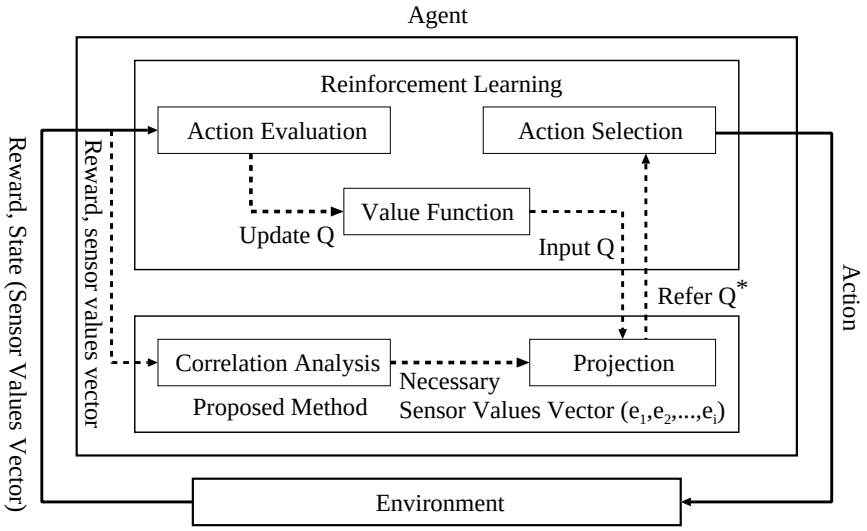


図 9: 具体的な学習機構

ルに相当するものである．すなわち，ロボットが自律的にセンサ情報を選択することを実現する機能である．Projection というモジュールは，Selective Use というモジュールに相当するものである．すなわち，ロボットが効率的に行動選択することを実現する機能である．

これらのモジュールは，強化学習の過程で次のように機能する．ここで，以下で説明する過程は，ロボットが $n(n > 0)$ 個のセンサを搭載していることを前提としてる．

1. エージェントは，現時刻 $T(T \geq 0)$ の状態 s_T を認識する．
2. 時刻 $T > 1$ のとき，センサ情報 $E_{1,T}, E_{2,T}, \ldots, E_{m,T}$ を選択する[9] ．
   (a) 時刻 T が $T > 0$ を満たすとき，現時刻 T までに標本した全てのセンサ情報 $E_{1,T}, E_{2,T}, \ldots, E_{n,T}$ について，報酬 R_T と相関分析する．このとき，(7) 式に従って，n 個の相関係数 $\rho_{1,T}, \rho_{2,T}, \ldots, \rho_{n,T}$ を算出する．ただし，(7) 式の分母が 0 となる場合は，$\rho_{i,T} = 0$ $(1 \leq i \leq n)$ とする．
   (b) n 個の相関係数 $\rho_{1,T}, \rho_{2,T}, \ldots, \rho_{n,T}$ について，その絶対値 $|\rho_{i,T}|(1 \leq i \leq n)$ が任意に設定した閾値 $\rho(\rho \geq 0)$ 以上であるかを判定する．ここで，$|\rho_{i,T}| \geq \rho$ であれば必要，$|\rho_{i,T}| < \rho$ であれば不要と選択する．このとき，以下では $m(0 \leq m \leq n)$ 個のセンサ情報 $E_{1,T}, E_{2,T}, \ldots, E_{m,T}$ が必要，$n - m$ 個のセンサ情報 $E_{m+1,T}, E_{m+2,T}, \ldots, E_{n,T}$ が不要と選択されたものとする．

[9]相関係数 ρ は，2 変量の標本が少なくとも 2 つ以上なければ求めることができない．

3. エージェントは，行動選択の方法に従い，行動 a_T を選択する．このとき，エージェントは次の場合分けに従って価値を参照する．

   - $T \leq 1$ ，$m = 0$，$m = n$ のいずれかを満たすとき，状態 s_T における任意の行動 a の価値 $Q(s_T, a)$ を参照する．
   - 上項の条件を満たさないとき，$E_{1,T}, E_{2,T}, \ldots, E_{m,T}$ により，(8) 式もしくは (9) 式に従って状態 s_T^* を再定義する．そして，状態 s_T^* における任意の行動 a の価値 $Q^*(s_T^*, a)$ を，(10) 式もしくは (11) 式に従って計算し，参照する．

4. エージェントは，次時刻 $T+1$ における状態 s_{T+1} を認識して，その状態に遷移する．また，このときに即時報酬 r_{T+1} を獲得する．さらに，センサ情報 $e_{1,T+1}, e_{2,T+1}, \cdots, e_{n,T+1}$ と報酬値 r_{T+1} を， 自律的選択に関する知識テーブルに追加する．

5. エージェントは，行動評価の方法に従い，状態 s_T にける行動 a_T の価値 $Q(s_T, a_T)$ を評価する．また，このときに，その価値 $Q(s_T, a_T)$ の評価回数 $N(s_T, a_T)$ を更新して，選択的利用に関する知識テーブルに記憶する．

6. 時刻 $T+1$ を時刻 T に改める．

7. タスクの実行を完了するまで，始めの手順に戻って繰り返す．

以上の方法により，本研究が提案する学習機構を適用したロボットは，自律的なセンサ情報の選択による状態認識と，それに基づいた効率的な行動選択できる．そして，そのロボットは，タスクを効率的に達成することができると考えられる．

5　実験

本章では，第 4 章で述べた提案手法の有効性を検証するための実験について述べる．そのために，本章で述べる 3 つ実験では，一般的な強化学習を適用したエージェントと，提案手法を適用したエージェントの性能を比較する．第 1 節と第 2 節では，これをシミュレーションによって検証する．これらの実験では，試行の終了条件が異なるタスクを用いて検証する．第 3 節では，シミュレーションで想定した環境とロボットを再現し，実際の実験によって検証する．この節では，第 1 節の実験のときと同様な試行の終了条件のタスクを用いて検証する．そして，それらの実験で得られた結果から，提案手法の有効性について考察する．

5.1　仮想環境における検証実験 1

5.1.1　実験目的

本実験は，提案手法の有効性を検証することを目的とする．そのために，提案手法を適用したエージェントと一般的な強化学習を適用したエージェントの 2 体を用意し，同一のタスクを実行させる．このような実験を，コンピュータ上でシミュレーションする．そして，その実験結果をもとにして，それぞれの性能を比較または考察する．これにより，提案手法の有効性を示す．

5.1.2　実験設定

本実験では，ロボットが存在する環境を，正方形状に配置した壁によって構成する．そして，その環境の中で，ロボットは「前方に存在する壁の近傍に到達する」というタスクを実行する．

はじめに，このタスクの概要を表現した図を図 10 に示す．図の通り，ロボットは正方形状に配置された壁に囲まれている．ロボットの前方に存在する壁は，図の壁 A に相当する．すなわち，ロボットがタスクを完了するためには，この壁 A の近傍（網掛けで描かれた領域）に到達すればよい．ロボットは，前方に存在する壁 A との距離を計測するセンサ 1 と，右方に存在する壁 B との距離を計測するセンサ 2 の 2 つを搭載している．ロボットは，これらの 2 つのセンサによって壁を認識する．ここで，ロボットがこのタスクを実行するためには，前方に存在する壁 A との距離を計測するセンサ 1 さえ搭載されていればよい．すなわち，センサ 1 がタスクを実行する上で必要であると言える．逆に，右方に存在する壁 B との距離を計測するセンサ 2 は，このタスクを実行する上で不要である．ロボットは，前後左右のいずれかの方向へ一定の距離だけ移動することができる．ただし，壁は十分な重量と強度

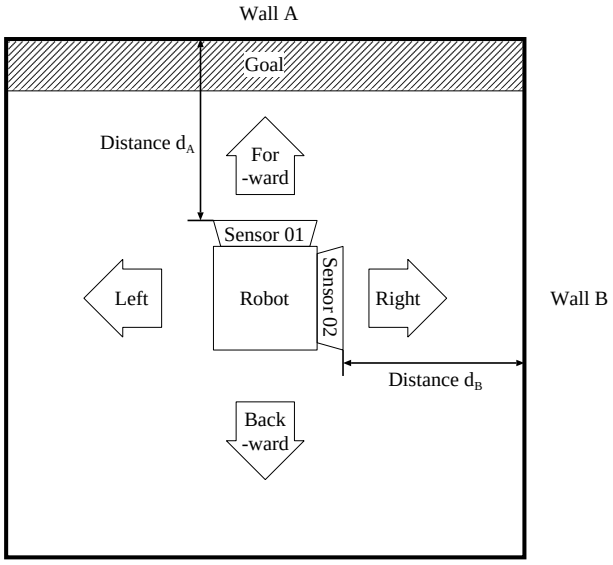


図 10: タスクの概要

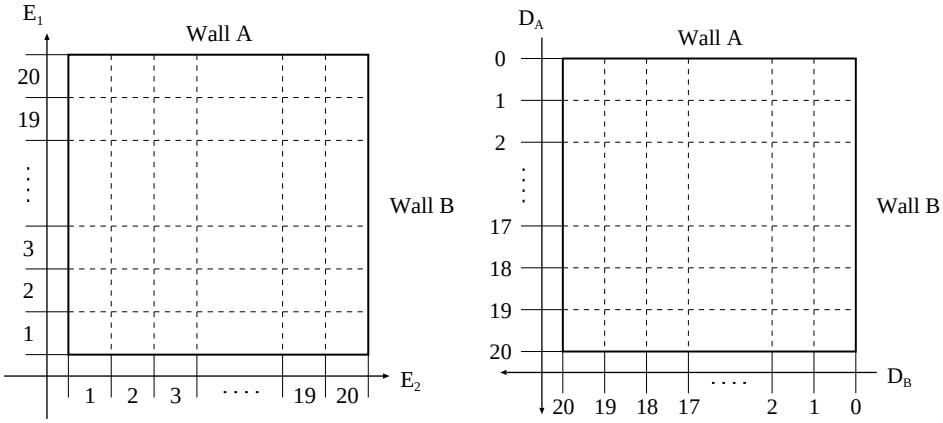
を持つ剛体とする．すなわち，壁はロボットに衝突されても移動することや壊れることはない．また，ロボットは壁を越えて移動することはできない．

これらのことを，強化学習について言い換えて述べる．すなわち，以下では強化学習を適用したロボットのことを指してエージェントと呼ぶ．エージェントは，前方に存在する壁 A との距離に対応するセンサ情報 E_1 と，右方に存在する壁との距離に対応するセンサ情報 E_2 によって状態を認識する．エージェントの状態をセンサ情報によって定義する前に，はじめに，センサ情報と壁との距離の関係について述べる．ここで，エージェントのセンサ情報 $e_1 \in E_1, e_2 \in E_2$ と距離 $d_A \in D_A, d_B \in D_B$ の関係を表現した図を図 11 に示す．また，センサ情報 E_1, E_2 と距離 D_A, D_B の関係を表す写像 f_A, f_B は次の通りである．

$$\begin{cases} f_A : D_A \mapsto E_1 \\ f_A(d_A) = \{e_1 \in E_1 | e_1 = \lceil 20 - d_A \rceil, 0 \leq d_A < 20\} \end{cases} \tag{12}$$

$$\begin{cases} f_B : D_B \mapsto E_2 \\ f_B(d_B) = \{e_2 \in E_2 | e_2 = \lceil 20 - d_B \rceil, 0 \leq d_B < 20\} \end{cases} \tag{13}$$

図の通り，センサ情報 $e_1 \in E_1, e_2 \in E_2$ は，それぞれの方向に正対している壁との距離 $d_A \in D_A, d_B \in D_B$ に比例した自然数である．これらの自然数は，壁との距離をある一定の長さに分割した区間に割り当てられており，エージェントが壁の近傍に存在しているときに最大値 20，エージェントが壁から最も遠い位置に存在しているときに最小値 1 をとる．



(a) センサ情報 E_1, E_2 が取り得る値　　(b) 距離 D_A, D_B が取り得る値

図 11: センサ情報 E_1, E_2 と距離 D_A, D_B の関係

このことを踏まえて，エージェントの状態をセンサ情報によって定義する．ここで，ある時刻 t におけるセンサ情報が $e_{1,t} \in E_1, e_{2,t} \in E_2$ であるとする．このとき，3.4 節の (3) 式に従って，ある時刻 t におけるエージェントの状態 s_t は，次式の通りに定義される．

$$s_t := \{(e_{1,t}, e_{2,t}) | e_{1,t} \in E_1, e_{2,t} \in E_2\} \tag{14}$$

エージェントは，上式に従って定義された状態に基づいて，行動を選択する．

次に，エージェントが存在する環境を，強化学習における状態に置き換えて説明する．ここで，図に示した環境を，強化学習における状態として表現した図を図 12 に示す．図の通り，エージェントが遷移可能な状態群 S を状態遷移図として表現すると，格子状のグラフのように描ける．本実験では，この格子状の状態群を 20×20 の大きさとする．エージェントが遷移可能な状態群は有限であるため，この環境は有限 MDP として見ることができる．図中の行動 a_1, a_2, a_3, a_4 は，それぞれ順に，前方移動，右方移動，後方移動，左方移動を意味している．エージェントは全ての状態において，前方移動，右方移動，後方移動，左方移動のいずれかの行動を選択可能である．エージェントの状態遷移は確定的であり，選択した行動の移動方向に隣接する状態へ必ず遷移できるものとする．ただし，壁に隣接する位置に相当する状態にエージェントが存在し，エージェントがそこで壁側の方向へ移動する行動を選択したときは，再び同じ状態に遷移するものとする．この制約は，エージェントが壁を越えて移動できないことを意味している．

さて，ここでエージェントに与える報酬について述べる．報酬は，タスクの進捗度に応じてエージェントに与えるものとする．このタスクでは，壁 A との距離 D_A がタスクの進捗度と相関があるものとする．ここで，ある時刻

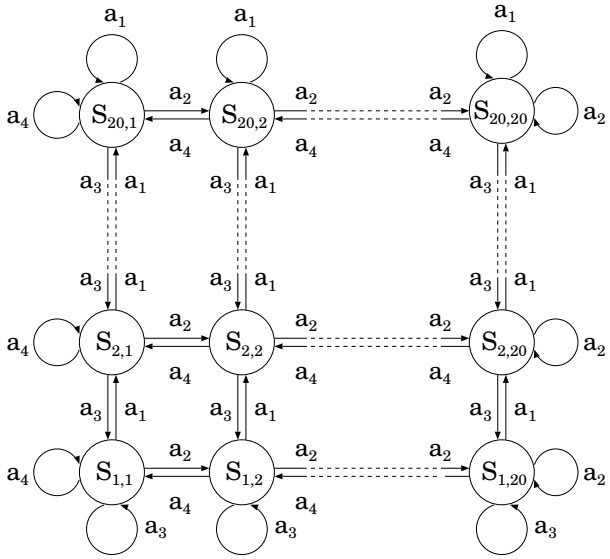


図 12: 環境の状態遷移図

t における壁 A との距離が $d_{A,t} \in D_A$ であるとする．このとき，エージェントに与えられる報酬 r_t は，次式の通りに定義される．

$$r_t = 20 - f_A(d_{A,t}) \quad (= 20 - e_{1,t}), \quad (1 \leq r_t \leq 20) \tag{15}$$

両エージェントには，1 時刻毎に上式で決定される報酬 r_t を与えられる．特に，この報酬は毎時刻に与えられる即時報酬であることに注意されたい．これにより，提案手法を適用したエージェントは強化学習によって行動価値 $Q(s_t, a_t)$ を評価すると同時に，現在のタスクの進捗度を知ることに繋がる．

提案手法を適用したエージェントは，上式の報酬 r_t とそれぞれのセンサ情報 $e_{1,t}, e_{2,t}$ について 1 時刻毎に相関分析して，相関係数 $\rho_{1,t}, \rho_{2,t}$ を求めることになる．本実験では，提案手法を適用したエージェントがタスクを実行する上で必要であると判別する相関係数の閾値 ρ は，0.8 とする[10]．

エージェントの行動選択には $\epsilon-$greedy 法，行動評価には Q 学習を適用する．探査的な行動を選択する確率 ϵ，ステップサイズ・パラメータ α，割引率 γ は後記する表 3 の通りである．

以上のような実験設定のもと，エージェントにタスクを実行させる．エージェントは，図 12 における状態 $s_{1,1}$ を初期状態として試行を開始する．このタスクの 1 試行の終了条件は，30000 回の行動選択とする．エージェントは，このような試行を 1000 回だけ実行するものとする．

最後に，以上の実験設定を要約して記述した表を以下に示す．

[10]統計学において，ある 2 変量の相関係数 0.8 以上であるとき，一般にそれらの変量の間には「強い相関がある」とされるからである

表 3: 実験設定の要約

項目	内容
1 試行の終了条件	30000 回の行動選択
総試行回数	1000
選択可能な行動群 A	{ 前方移動，右方移動，後方移動，左方移動 }
遷移可能な状態数 $\|S\|$	20×20
行動選択	$\epsilon-$greedy
探査的な行動を選択する確率 ϵ	0.05
行動評価	Q 学習
ステップサイズ・パラメータ α	0.5
割引率 γ	0.5
行動価値 Q の初期値	0.0
相関係数の閾値 ρ	0.8

5.1.3 実験結果

本小節では，前小節で述べた設定をもとに実験した結果について述べる．試行の終了条件はある一定の行動回数である．そのため，エージェントが 1 試行の間により多くの報酬値を獲得すれば良いといえる．

そこで，本実験では，両エージェントがある時刻 t のときに獲得した報酬値の試行回数毎の和を，タスクの総試行回数で割ることで，1 試行間においてエージェントが 1 行動毎に獲得した報酬の平均値の推移を算出した．また，それと同様に計算して，1 試行間におけるセンサ情報 $e_{1,t} \in E_1, e_{2,t} \in E_2$ と報酬 $r_t \in R$ の間の相関係数 $\rho_{1,t}, \rho_{2,t}$ の平均値の推移も算出した．

はじめに，両エージェントが 1 行動ごとに獲得した報酬の平均値の推移を比較するグラフを図 13 に示す．本実験のタスクでは，報酬 r_t の最大値は 20 である．したがって，その平均値が 20 に近ければ近いほど良い結果と言える．図の通り，提案手法を適用したエージェントが 1 行動ごとに獲得した報酬値は，一般的な強化学習を適用したエージェントのものよりも，1 試行における時刻の全体において上回っている．これは，提案手法を適用したエージェントが，一般的な強化学習を適用したエージェントよりも早くに行動を学習していることを意味する．したがって，提案手法を適用したエージェントの方が良い結果を示していると言える．

次に，1 試行間におけるセンサ情報 E_1, E_2 と報酬 R の間の相関係数 $\rho_{1,t}, \rho_{2,t}$ の平均値の推移を比較するグラフを図 14 に示す．図の通り，壁 A との距離に対応するセンサ情報 $e_{1,t} \in E_1$ と報酬値 $r_t \in R$ との相関係数 $\rho_{1,t}$ が，直ちに閾値 $\rho = 0.8$ を超えて，1.0 に収束していることが分かる．また，壁 B

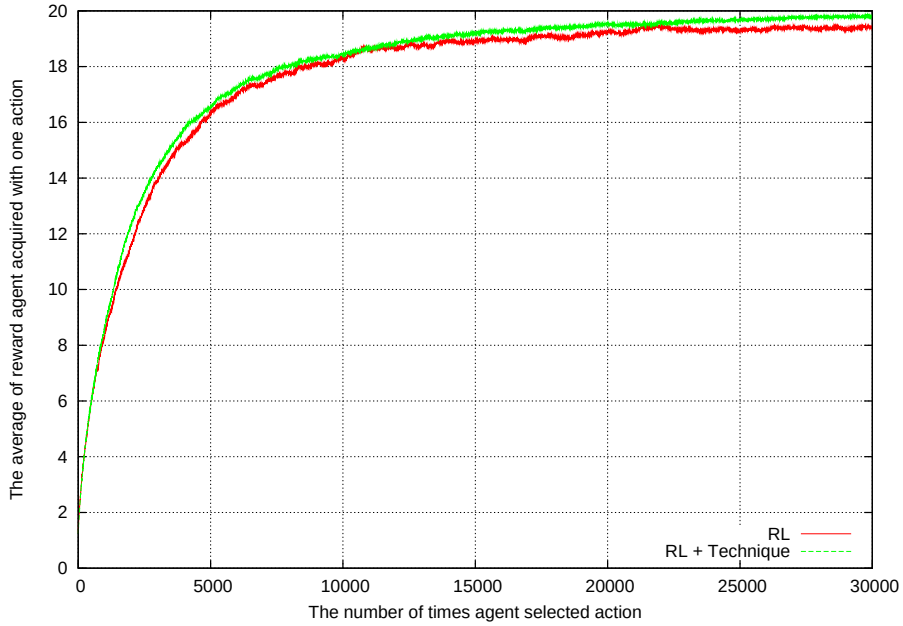


図 13: 報酬値 r_t の平均値の推移（$\epsilon = 0.05$ の時）

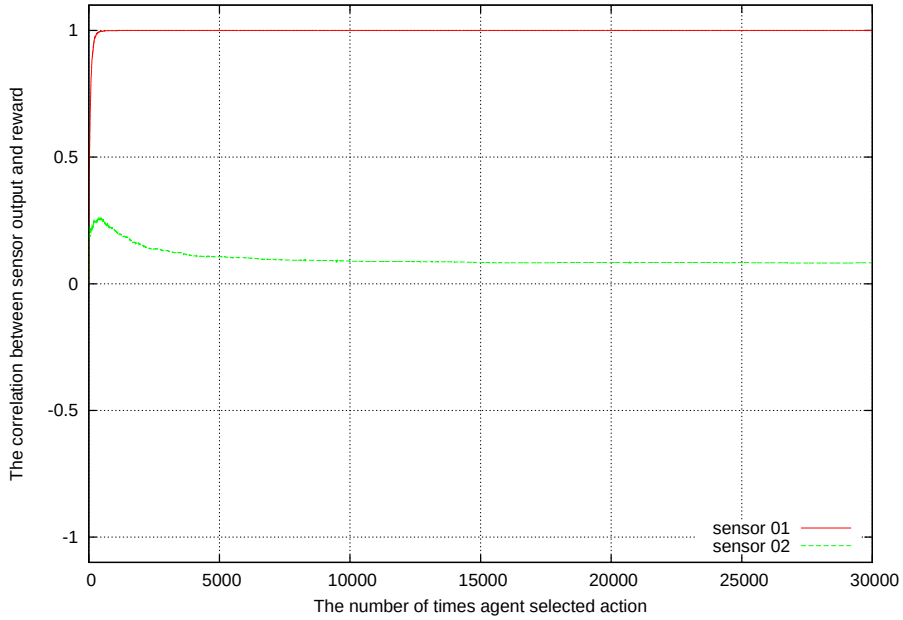


図 14: 相関係数 $\rho_{1,t}, \rho_{2,t}$ の平均値の推移（$\epsilon = 0.05$ の時）

との距離に対応するセンサ情報 $e_{2,t} \in E_2$ とその相関係数 $\rho_{2,t}$ が，0.1 程度に収束していることが分かる．これは，提案手法を適用したエージェントが，1 試行の初期からセンサ情報 $e_{1,t} \in E_1$ のみで定義される状態にもとづいて行動を選択していたことを意味する．

ここで，本実験では，$\epsilon-$greedy の探査的な行動を選択する確率 ϵ を 0.05 とした．補足として，他の実験設定は全く同一にして，探査的な行動を選択する確率 ϵ のみを $0.1, 0.2$ とした場合で実験した．それらの場合のときの実験結果も図 15, 図 16, 図 17, 図 18, に示す．これらの図の通り，探査的な行動を選択する確率 ϵ を変動させても，提案手法を適用したエージェントが一般に優位であることが分かる．

5.1.4 考察

本小節では，前小節で示した実験結果を考察する．はじめに，一般的な強化学習を適用したエージェントの学習の流れを表現した図を図 19 に示す．図の太い実線で引かれた枠は，エージェントを囲む正方形状の壁である．また，細線で引かれたマスは，強化学習における状態である．図の上部にある網掛けで描かれたマスは，その行動価値が学習によって真の値に収束している状態を表現している．図の実線の矢印は，貪欲な手として選択した行動による移動，破線の矢印は探査的な手として選択した行動による移動を表現している．

さて，はじめに，一般的な強化学習を適用したエージェントの初期の学習の流れについて考える．図の通り，エージェントは，壁 A，壁 B から最も遠い状態を初期状態としてタスクを開始する．この初期状態は，図の左下の状態に相当する．そして，エージェントは未探索な状態へ遷移を繰り返しながら，目標状態である壁 A の近傍に存在する状態へ少しずつ近付いていく．ここで，図の「Problem 1」で表現したように，エージェントは幾つかの同じ状態行動対をループすることがある．これは，全ての状態において即時報酬が与えられることと，エージェントが $\epsilon-$greedy によって確率 $1-\epsilon$ で貪欲な手として行動を選択することが原因である．これにより，エージェントが確率 ϵ で探査的な手として行動を選択しない限り，延々とその状態行動対のループを繰り返すことになる．この「Problem 1」のような問題は，十分な時刻だけエージェントが学習していないとき，あらゆる状態行動対で起こり得る．これが，一般的な強化学習を適用したエージェントの学習を遅らせる要因のひとつである．

次に，一般的な強化学習を適用したエージェントの中期の学習の流れについて考える．タスクをある一定の時刻だけ実行すると，エージェントは目標状態へ到達する．そして，エージェントは貪欲な手と探査的な手を繰り返すことで，目標状態とその近傍の状態の行動価値を収束させていく．これにより，図 19 のように網掛けの領域が増えていくのである．ここで，図の「Problem 2」で表現したように，エージェントが探索済みの目標状態とその近傍の状態

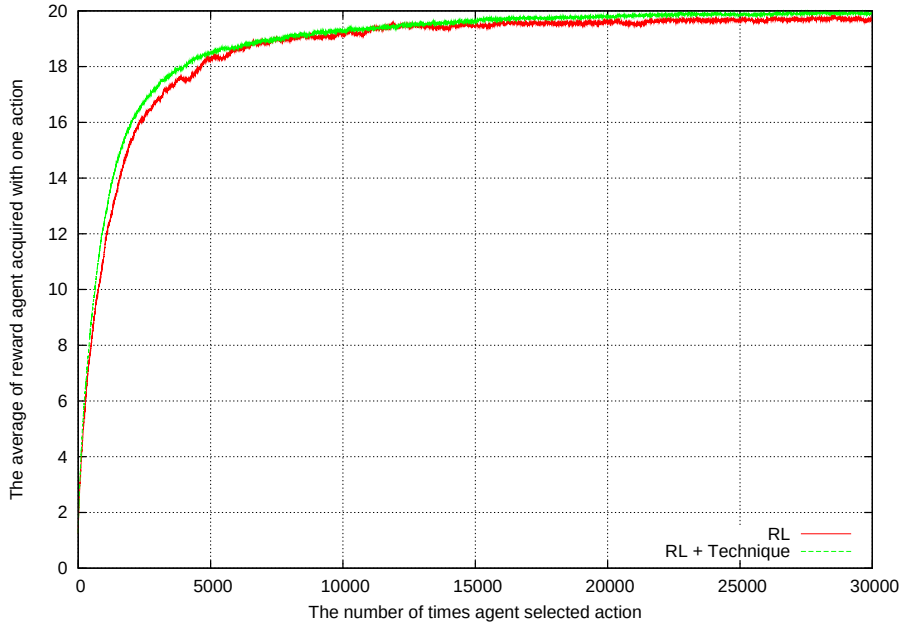


図 15: 報酬値 r_t の平均値の推移（$\epsilon = 0.10$ の時）

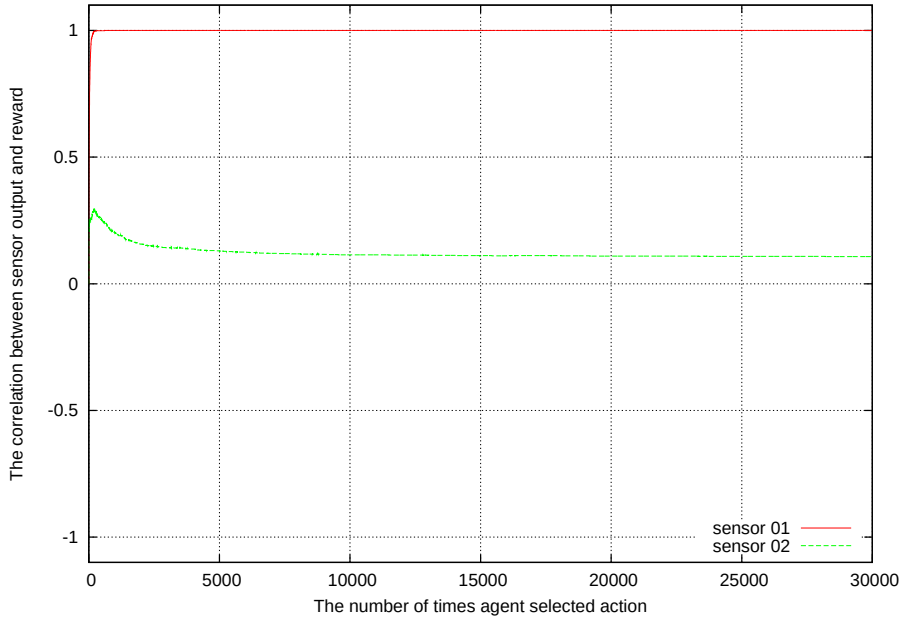


図 16: 相関係数 $\rho_{1,t}, \rho_{2,t}$ の平均値の推移（$\epsilon = 0.10$ の時）

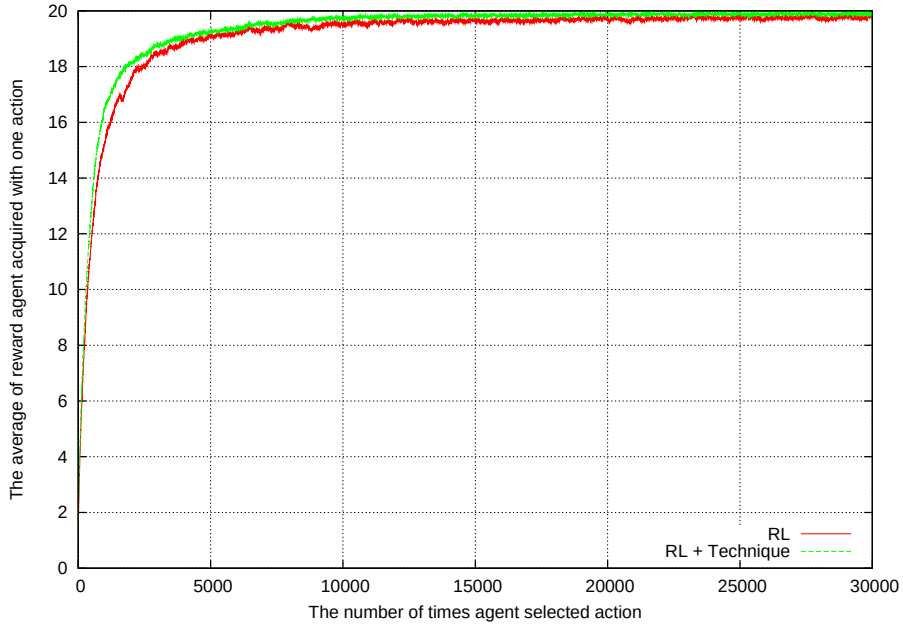


図 17: 報酬値 r_t の平均値の推移（$\epsilon = 0.20$ の時）

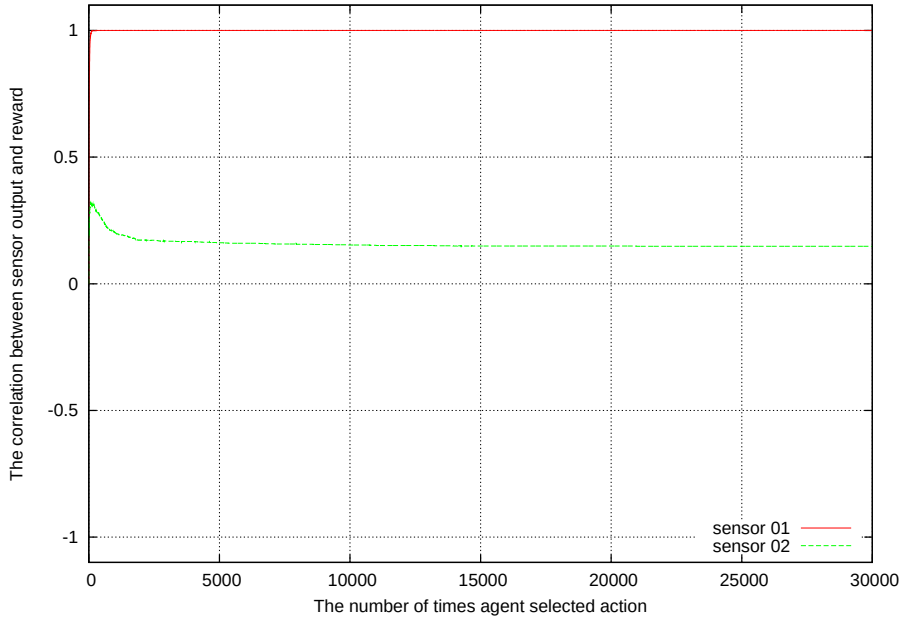


図 18: 相関係数 $\rho_{1,t}, \rho_{2,t}$ の平均値の推移（$\epsilon = 0.20$ の時）

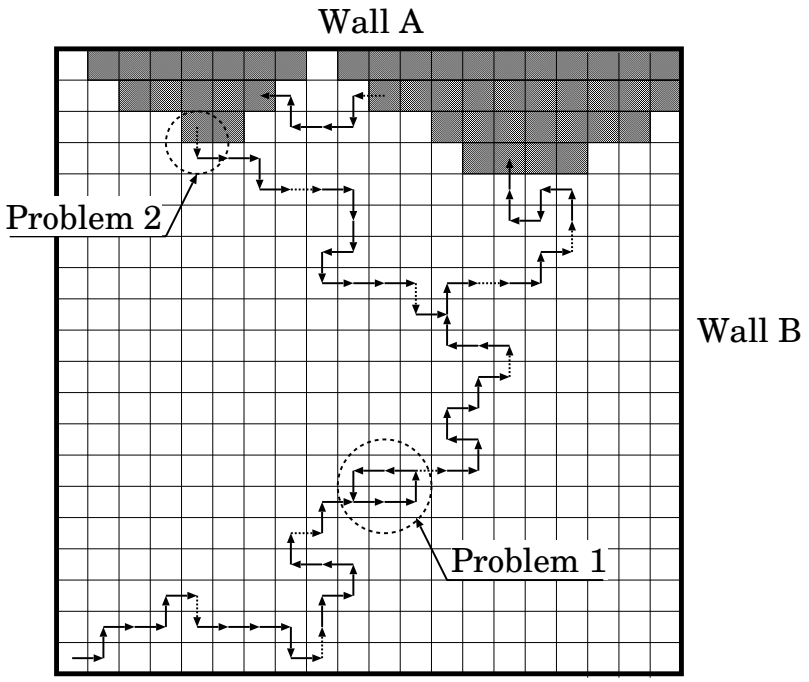
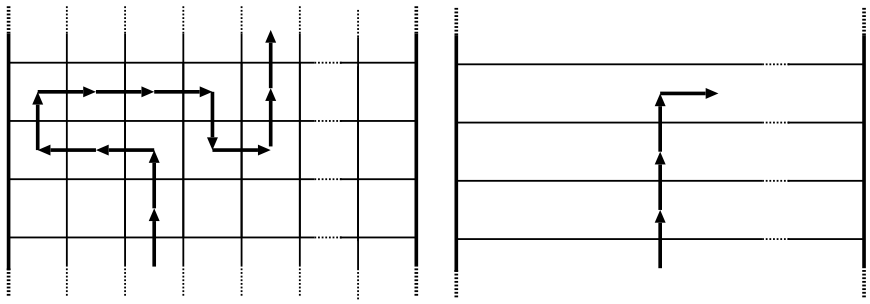


図 19: 一般的な強化学習を適用したエージェントの学習の流れ

の群の端で，探査的な手として行動を選択することがある．そして，このとき，エージェントは未探索な状態へ遷移することがある．未探索な状態では全ての行動価値が等しいから，そこで貪欲な手として行動を選択するとしても，エージェントはランダムに行動を選択するしかない．このとき，エージェントはまた新たに未探索な状態へ遷移することがある．こうして，エージェントは未探索な状態群へ迷い込むことがあるのである．その後，エージェントはしばらくして探索済みの状態群へ辿り着き，そこで獲得した行動によって再び目標状態へ戻る．この「Problem 2」のような問題は，十分な時刻だけエージェントが学習していないとき，1 試行において 2,3 回ほど起こり得る．これもまた，一般的な強化学習を適用したエージェントの学習を遅らせる要因のひとつである．

一般的な強化学習を適用したエージェントの場合に起こるこれら 2 つの問題は，全ての試行のどの時刻においても起こり得るものである．したがって，試行の全体に渡って，一般的な強化学習を適用したエージェントが獲得する報酬の平均値を引き下げることになるのである．

これらのことを踏まえて，次に，提案手法を適用したエージェントの学習の流れについて考える．はじめに，提案手法を適用したエージェントの学習の流れを表現した図を図 20 に示す．図 20 の見方は，図 19 と同様であるため，その説明は割愛する．



(a) $\rho_{1,t} < \rho$ の時の状態認識と行動　　(b) $\rho_{1,t} \geq \rho$ の時の状態認識と行動

図 20: 提案手法を適用したエージェントの学習の流れ

提案手法を適用したエージェントが試行を開始した直後は，一般的な強化学習を適用したエージェントと同様に状態を認識して学習する (図 20(a))．すなわち，このときに両エージェントの行動選択に違いはない．しかし，このタスクにおける環境で幾つかの状態を遷移すると，壁 A との距離に対応するセンサ情報 $e_1 \in E_1$ と報酬 $r \in R$ の相関係数 $\rho_{1,t}$ は，直ちに $\rho_{1,t} \geq \rho$ を満たす．すなわち，提案手法を適用したエージェントは，壁 A との距離に対応するセンサ情報 $e_1 \in E_1$ のみで定義される状態 s^* を認識し，その行動価値 Q^* を参照して行動を選択する (図 20(b))．

さて，ここで，先に述べた「Problem 1」および「Problem 2」のような問題と，提案手法を適用したエージェントの関係に着目する．

はじめに,「Problem 1」について考える．提案手法を適用したエージェントは，提案手法が機能しているとき，図 20(b) のように低次元化した状態を認識する．そのため，本来の状態行動対 (s, a) の個数と比べて，それよりも少ない個数の (s^*, a) について学習する．これより,「Problem 1」のように同じ状態行動対のループが発生する確率がその分だけ低くなるのである．これが，提案手法を適用したエージェントが，一般の強化学習を適用したエージェントよりも，報酬の平均値で上回る要因のひとつである．

次に,「Problem 2」について考える．エージェントが探索済みの目標状態とその近傍の状態の群の端で，探査的な手として行動を選択するとき，エージェントは未探索な状態へ遷移することがあると述べた．ここで，提案手法を適用したエージェントは，第 4.2 節で記述した (10) 式に従って行動価値 $Q^*(s_t^*, a)$ を計算する．すなわち，提案手法を適用したエージェントは，本来はその時点で未探索な状態に遷移したとしても，同じセンサ情報 $\exists e_1 \in E_1$ で定義される他の状態の行動価値 Q を低次元化的に写像した行動価値 $Q^*(s_t^*, a)$ を参照する．そのため，提案手法を適用したエージェントは，過去の経験から貪欲な手として行動を明確に選択することができる．これにより，このエージェントは未探索な状態群へ迷い込まないのである．これが，提案手法を適用し

たエージェントが，一般の強化学習を適用したエージェントよりも，報酬の平均値で上回る要因のひとつである．

提案手法を適用したエージェントの場合，全ての試行のほとんどの時刻において提案手法が機能していたため，2 つの問題は起こらなかったようである．したがって，試行の全体に渡って，提案手法を適用したエージェントが獲得する報酬の平均値が，一般の強化学習を適用したエージェントのものよりも上回ることになったのである．

5.2 仮想環境における検証実験 2

5.2.1 実験の目的と設定

本研究では，試行の終了条件以外を全く同じ実験設定で，同様に実験した．そのため，本節での実験設定の説明については割愛する．また，この実験の目的は前節のものと同じである．

本実験の 1 試行の終了条件は，目標状態 $s_{20,1}, s_{20,2}, \ldots, s_{20,20}$ のいずれかへの遷移である．この実験も，コンピュータ上でシミュレーションする．

本節の実験設定を要約して記述した表を以下に示す．

表 4: 実験設定の要約

項目	内容
1 試行の終了条件	目標状態 $s_{20,1}, s_{20,2}, \ldots, s_{20,20}$ のいずれかへの遷移
総試行回数	10000
選択可能な行動群 A	{ 前方移動，右方移動，後方移動，左方移動 }
遷移可能な状態数 $\|S\|$	20×20
行動選択	$\epsilon-$greedy
探査的な行動を選択する確率 ϵ	0.05
行動評価	Q 学習
ステップサイズ・パラメータ α	0.5
割引率 γ	0.5
行動価値 Q の初期値	0.0
相関係数の閾値 ρ	0.8

5.2.2 実験結果

本実験のタスクはエピソード型である．そのため，エージェントが 1 試行の間により少ない行動回数で目標状態に遷移すれば良いといえる．

そこで，本実験では，両エージェントが 1 試行当たりに費やした行動回数を記録した．両エージェントが 1 試行当たりに費やした行動回数を比較するグラフを図 21, 図 22 に示す．

本実験のタスクでは，エージェントはタスクを達成するために少なくとも 19 回の行動を要する．すなわち，ある試行における行動回数が低く，19 に近ければ近いほど良い結果といえる．図 21 の通り，一般的な強化学習を適用したエージェントが要した行動回数が，500 試行目の前後，2500 試行目の付近，5000 試行目の前，6500 試行目の後，7000 試行目の前で，インパルス状に増えていることが分かる．また，図 22 の通り，一般的な強化学習を適用したエージェントは，試行の全体に渡って，提案手法を適用したエージェントよりも多く行動回数を要していることが分かる．

したがって，提案手法を適用したエージェントの方が良い結果を示していると言える．

5.2.3 考察

一般的な強化学習を適用したエージェントは，試行の全体に渡って，提案手法を適用したエージェントよりも多く行動回数を要している．これは，前節の実験結果の考察で述べた「Problem 1」が原因であると考えられる．すなわち，一般的な強化学習を適用したエージェントは，未探索な状態へ遷移を繰り返していく過程で，幾つかの同じ状態行動対をループすることが原因である．さらに，Q 学習によって目標状態から順に価値がフィードバックしてくることを考えると，本実験のタスクはエピソード型であるから，その分だけそのようなループが解消されるのに多くの時間を要していることが考えられる．これらのことから，一般的な強化学習を適用したエージェントは，試行の全体に渡って多くの行動回数を要していると予測される．

また，一般的な強化学習を適用したエージェントが要した行動回数が，幾つかの試行のときにインパルス状に増えていることに着目する．これは，前節の実験結果の考察で述べた「Problem 2」が原因であると考えられる．すなわち，エージェントが探索済みの状態群の端で探査的な手として行動を選択したときに，未探索な状態へ続けて遷移してしまうことが原因である．さらに，この問題も同様に，本実験のタスクはエピソード型であるから，その分だけ未探索な状態群における行動価値の収束が遅れがあることや，未探索の状態が残存していることが考えられる．これらのことから，一般的な強化学習を適用したエージェントが要した行動回数が，しばしばインパルス状に増えてしまったと予測される．

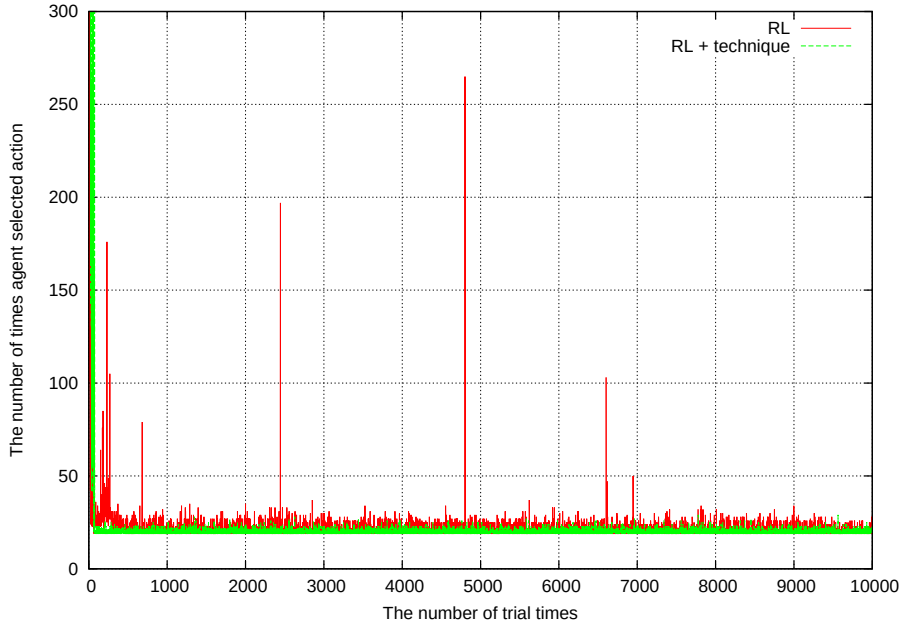


図 21: 両エージェントが 1 試行当たりに費やした行動回数

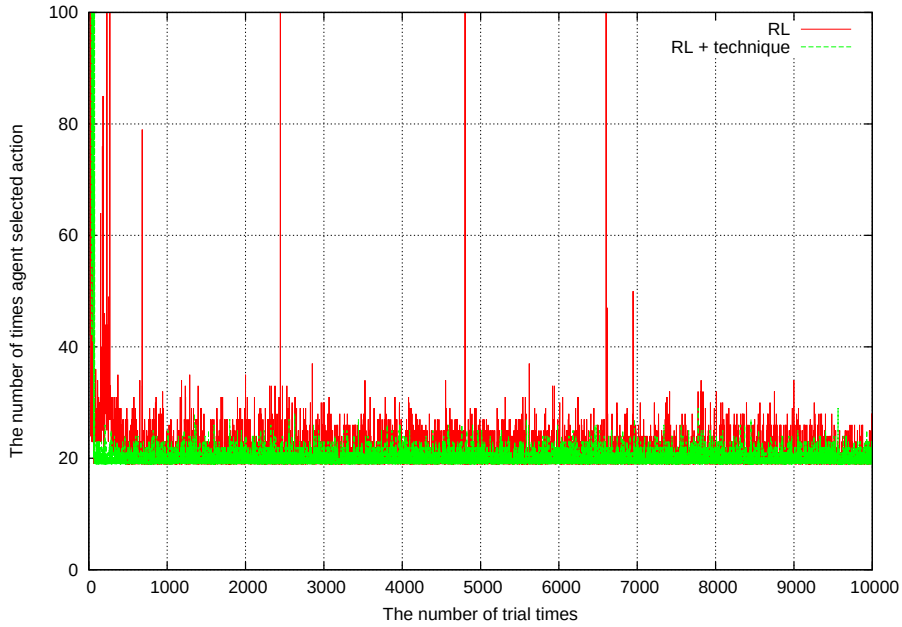


図 22: 上図の縦軸を区間 $[0, 100]$ に拡大したもの

5.3 実環境における検証実験

5.3.1 実験目的

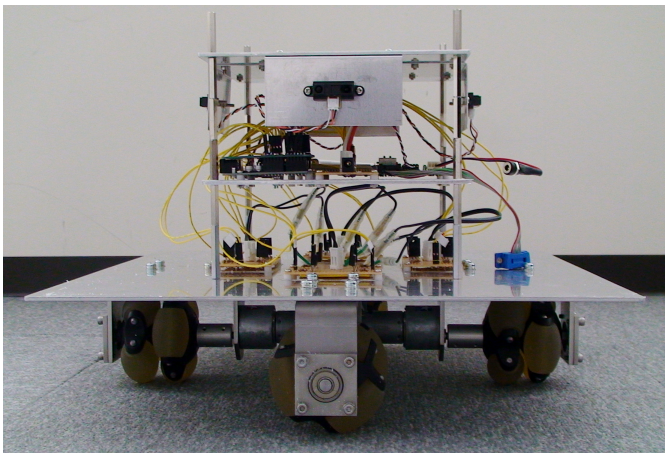
本実験は，提案手法の有効性を検証することを目的とする．そのために，提案手法を適用したエージェントと一般的な強化学習を適用したエージェントの 2 体を用意し，同一のタスクを実行させる．本実験では，実際のロボットを使用する．そして，その実験結果をもとにして，それぞれの性能を比較または考察する．これにより，提案手法の有効性を示す．

5.3.2 ロボットの構成とセンサの関係

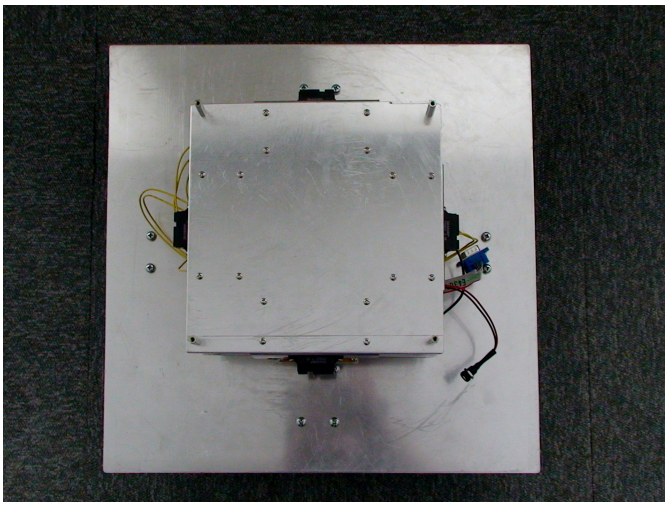
本実験で使用するロボットの構成について説明する．はじめに，ロボットを撮影した写真を図 23 に示す．ロボットの大きさは，高さ 262 (mm)，幅 400 (mm)，奥行き 400 (mm) である．ロボットの投影図および組立図については，本論文の最後に添付した付録を参考にされたい．

学習のプログラムを実行するプラットフォームには，Armadillo-300 を採用している．Armadillo-300 は，強化学習における行動や状態認識に伴い，Arduino UNO にモータ駆動の指令やセンサ情報の要求をする．Armadillo-300 と Arduino UNO は，AGB65-232C を介してシリアル通信によって双方向にデータ転送する．AGB65-232C は，シリアル通信の信号レベルを変換するインターフェイス回路である．Armadillo-300 は RS232C（PC）レベルの信号を出力し，Arduino UNO は TTL（5.0V マイコン）レベルの信号を出力する．AGB65-232C は，これらの信号のレベルを互換するためのものである．Arduino UNO がモータを駆動させるときは，H ブリッジ回路に制御信号を送る．H ブリッジ回路は，2 値 $\times$ 2 本の制御信号の組み合わせにより，DC モータを任意の方向に回転させることができる．これにより，ロボットは前後左右への移動動作と静止動作を任意に選択することができる．また，Arduino UNO がセンサ情報を取得するときは，赤外線センサ GP2Y0A21YK0F が出力する信号電圧を A/D コンバータで変換する．この A/D コンバータは，区間 $[0.0, 5.0]$ (V) の電圧を，1024 の数だけ均一に量子化し，区間 $[0, 1023]$ の整数値に変換する (図 24)．これにより，ロボットはセンサ情報によって壁からの距離を認識することができる．ロボットの回路図についても同様に，本論文の最後に添付した付録を参考にされたい．

ここで，GP2Y0A21YK0F の電圧-距離の特性関係を示すグラフを図 25(a) に示す．このグラフは，参考文献 [26] から抜粋したものである．図 25(a) を見ると，0(cm) から 5(cm) 強未満の範囲，およびそれ以降から 80(cm) 以下の範囲で，電圧の取り得る値が重複していることが分かる．すなわち，GP2Y0A21YK0F が出力する電圧信号から，距離を一意に特定することができない．これより，GP2Y0A21YK0F は 5(cm) 強未満の距離を計測するのは不



(a) ロボットの正面



(b) ロボットの上面

図 23: ロボットの写真

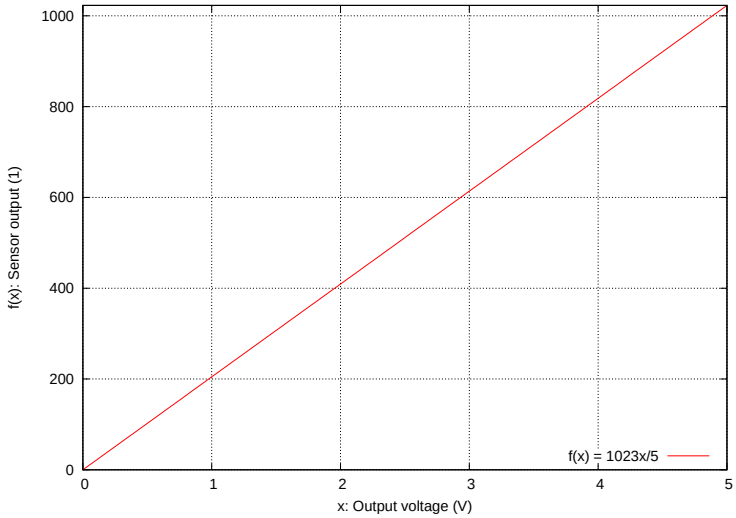


図 24: Arduino UNO の A/D コンバータ

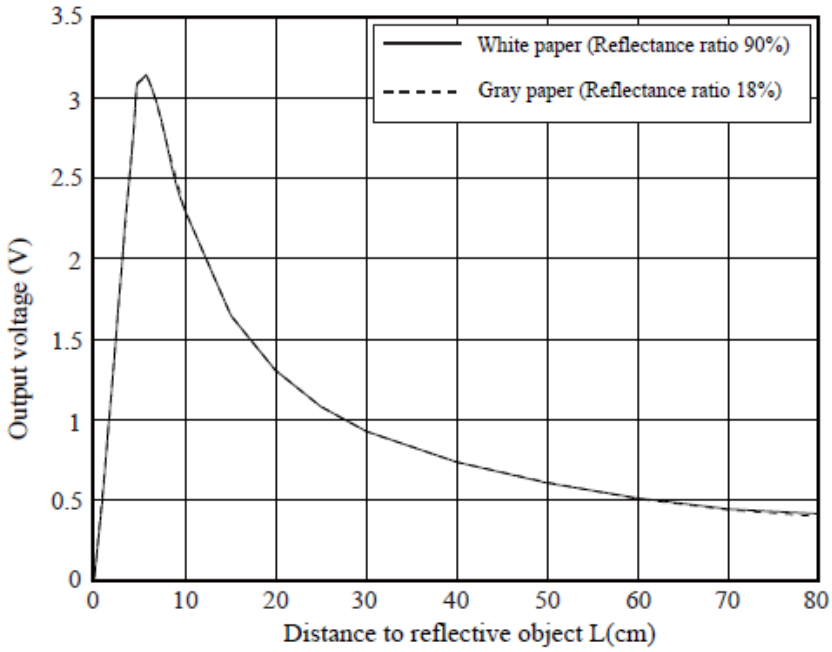
適等であると言える．そこで，付録の第 2 節の図 52 の通り，GP2Y0A21YK0F をロボットの第 3 階層の端よりも 10(cm) だけ奥に取り付けた．これにより，GP2Y0A21YK0F が距離を計測する上で問題となる範囲を十分に包含する距離の区間 $[0, 10)$ は，構造的に計測されないようにした．

また，この赤外線センサ GP2Y0A21YK0F の電圧-距離の特性関係は，資料である参考文献 [26] の中で，その関係式を明示されていない．あくまで，図 25(a) のような概形を示したグラフだけの記載である．そこで，本研究ではその関係式を $f(x) = \frac{32}{(x+4)}$ で近似する (図 25(b))．すなわち，$f(x)$ が電圧，x が距離に相当する．本実験で使用するロボットは，この近似式に従って，GP2Y0A21YK0F が計測する距離を求めるものとする．

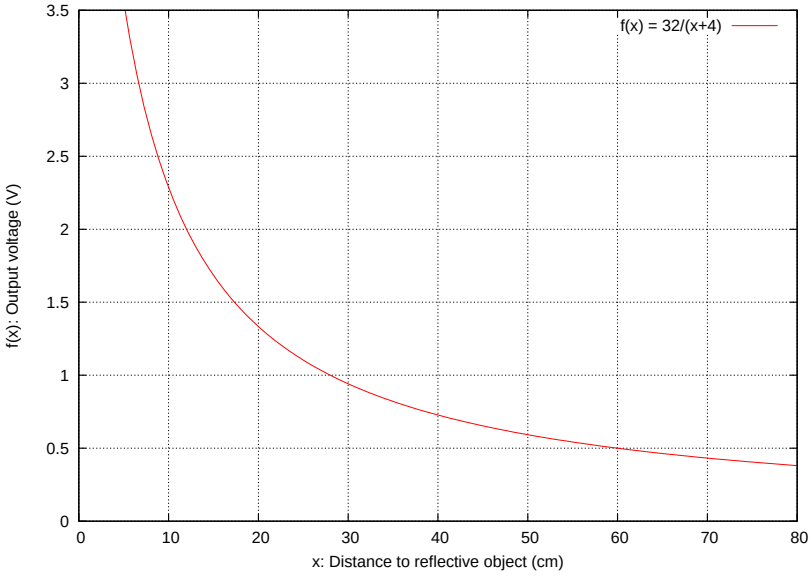
5.3.3 実験設定

本実験では，ロボットが存在する環境を正方形状に配置した壁によって構成する．そして，その環境の中で，ロボットは「前方に存在する壁の近傍に到達する」というタスクを実行する．すなわち，本実験におけるタスクとその概要は，第 1 節のものと同様である．

はじめに，ロボットがタスクを実行する環境を撮影した写真を図 26 に示す．環境を構成する壁には，ダウ化工株式会社が販売する「スタイロフォーム IB」という製品を使用する．本実験では，厚さが 20mm の スタイロフォーム IB の大きさを，高さ 300mm，幅 1100mm に加工して，これを 1 枚の壁とする．この壁を 4 枚作り，それらを正方形状に配置することで，実験環境



(a) 赤外線センサ GP2Y0A21YK0F の電圧-距離の特性関係



(b) 近似式 $f(x) = \frac{32}{(x+4)}$

図 25: 赤外線センサ GP2Y0A21YK0F の電圧-距離の特性関係とその近似式

を実現した．ちなみに，これらの壁は図 27 のように，蝶番とネジによって互いに連結した．これにより，壁は支えを必要とせずに，壁全体が床に接地することで直立することができる．次に，この環境とロボットの大きさの関係

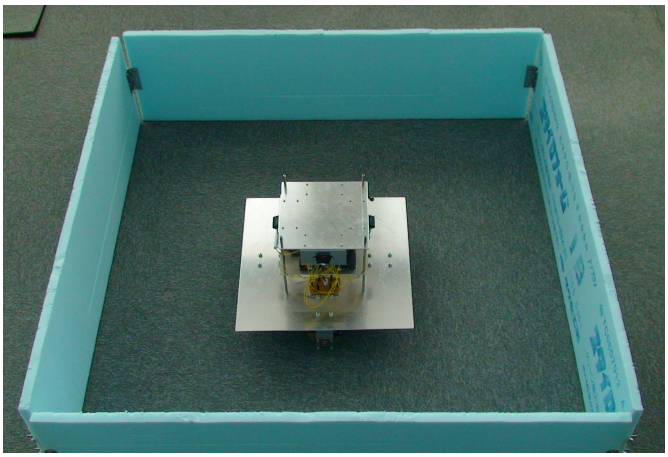
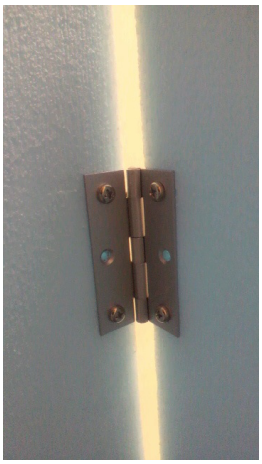


図 26: ロボットがタスクを実行する環境

(a) 外側の連結

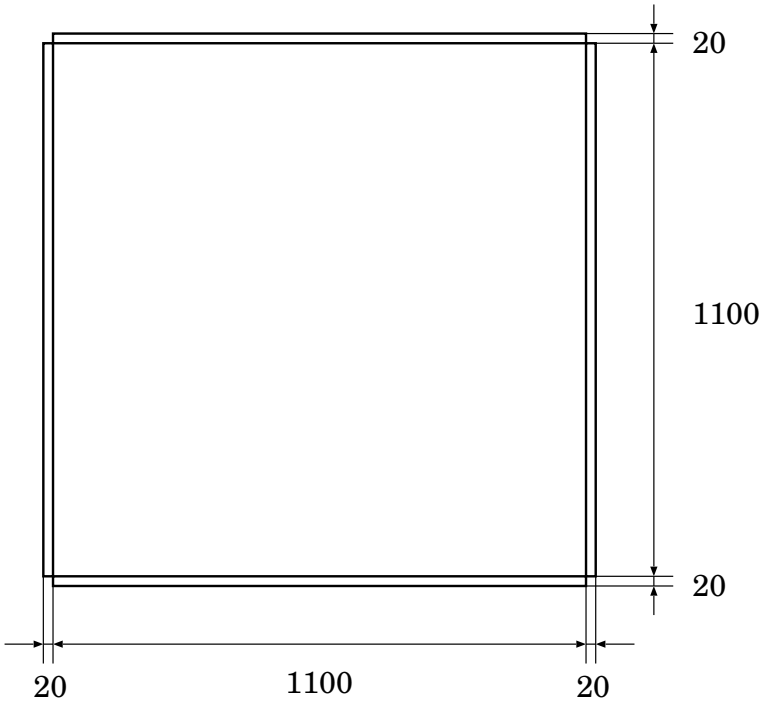


(b) 内側の連結

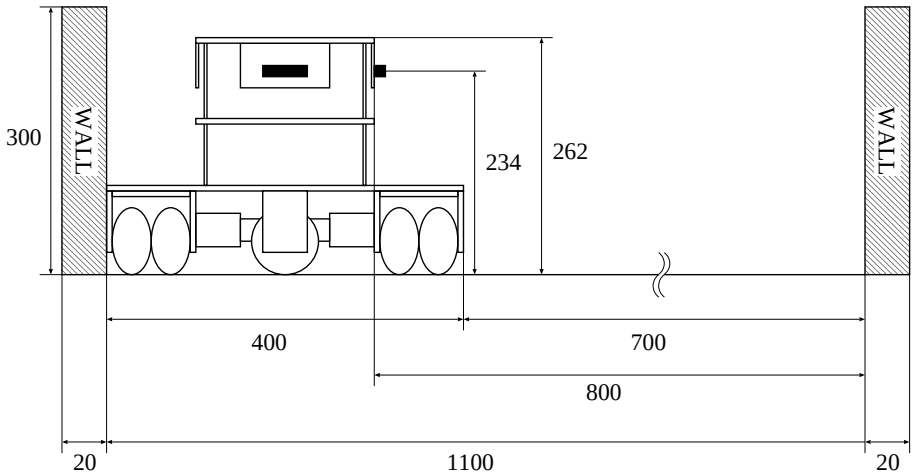
図 27: スタイロフォーム IB を使用した壁とその連結

を表現したものを図 28 に示す．図の通り，この環境でロボットが移動できる範囲は 1100 mm × 1100 mm である．赤外線センサ GP2Y0A21YK0F を搭載する位置の高さが 234 mm 付近であるのに対して，壁の高さは 300 mm である．これにより， GP2Y0A21YK0F を搭載する位置の傾きの誤差に依らず，GP2Y0A21YK0F は壁を十分に認識することができる．このような環境の中で，ロボットはタスクを実行する．

次に，エージェントの強化学習における状態認識の設定について述べる．こ



(a) 上面から見た環境の寸法



(b) 側面から見た環境とロボットの寸法

図 28: 環境とロボットの大きさの関係

こで，エージェンとは強化学習を適用したロボットであることに注意されたい．本実験では，エージェントは実環境の状態を認識する．そのため，第 3.4 節で述べた通り，式 (3) のようにセンサ情報の列によって状態を定義するのではなく，式 (6) のようにそれを離散化した状態変数の列によって状態を定義することが望ましいと考えられる．そこで，まずは本研究が考えるエージェントの状態認識の設定を表現したものを図 29 に示す．図の通り，エージェン

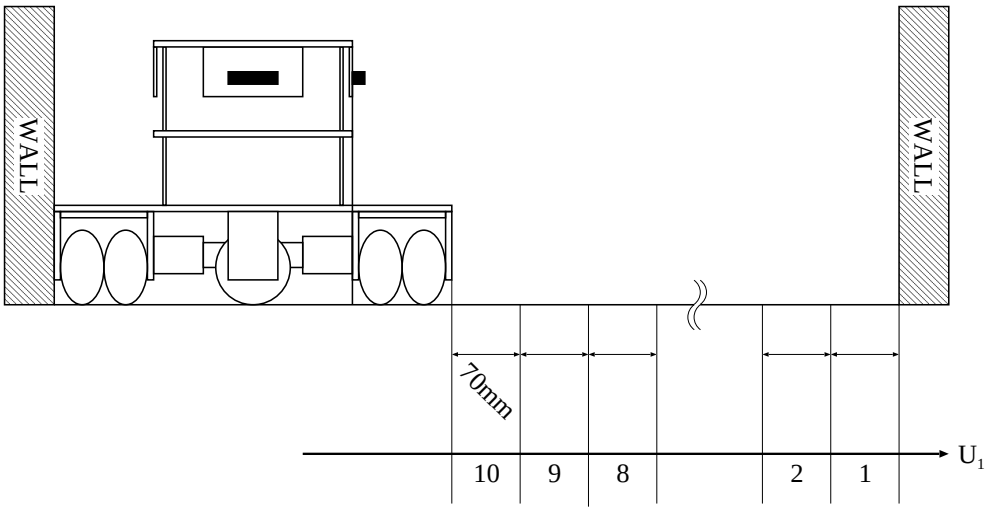


図 29: エージェントの状態認識の設定

トは壁から 70 mm の間隔で等しく状態を認識するように設定する．このとき，離散化した距離の区間には，壁から順に自然数を昇順に割り振る．これらの自然数とその集合が，状態変数 $u \in U$ に相当する．ここで，ある距離とその集合を d, D，ある状態変数とその集合を u, U とおく．このとき，ある状態変数 u は，ある距離 d から次の写像 g' によってマッピングされるものとする．

$$\begin{cases} g' : D \mapsto U \\ g'(d) = \{u = i | (i-1) \times 70 \leq d < i \times 70, i \in N\} \end{cases} \tag{16}$$

紙面鉛直向きの方向の状態認識についても同様である．しかしながら，ノイズによる計測誤差によって，センサは壁からの距離を 700 mm 以上と測定することがある．すなわち，エージェントが遷移可能な状態群 S の個数 $|S|$ は，設定上では 10×10 (=100) であるが，実際には 11×11 (=121) となる．エージェントが遷移可能な状態群は有限であるため，この環境は有限 MDP として見ることができる．しかしながら，この環境における状態遷移は非決定的—確率的—であることに注意されたい．これは，強化学習における状態認識時において，センサ情報にノイズが混入することによって不完全認識が

発生するためである．また，強化学習における行動時において，床が不整地であるためにモータが空転するためである．

さて，先に述べた状態認識の設定を実現するためには，センサ情報 $e \in E$ と壁からの距離 $d \in D$ を対応させなければならない．すなわち，あるセンサ情報 $e \in E$ からある電圧 $v \in V$ への写像 f_1，およびある電圧 $v \in V$ から壁からの距離 $d \in D$ への写像 f_2 を考えなければならない．

はじめに，あるセンサ情報 $e \in E$ からある電圧 $v \in V$ への写像 f_1 を考える．この写像 f_1 は，前小節 5.3.2 で述べた Arduino UNO の A/D コンバータの関係 (図 24) より，次の通りに定義できる．

$$\begin{cases} f_1 : E \mapsto V \\ f_1(e) = \{v \in V | v = \frac{5}{1023} \times e, 0 \leq e \leq 1023\} \end{cases} \tag{17}$$

次に，ある電圧 $v \in V$ から壁からの距離 $d \in D$ への写像 f_2 を考える．この写像 f_2 は，前小節 5.3.2 で述べた赤外線センサ GP2Y0A21YK0F の電圧-距離の特性関係とその近似式より，次の通りに定義できる．

$$\begin{cases} f_2 : V \mapsto D \\ f_2(d) = \{d \in D | d = \frac{32}{v} - 4, 0 < v \leq 5\} \end{cases} \tag{18}$$

ただし，図 25(a) の通り，赤外線センサ GP2Y0A21YK0F は 5(V) まで電圧を出力しないことに注意されたい．

したがって，これらの写像 g', f_1, f_2 の合成写像 $f_1 \circ f_2 \circ g'(e)$ によって，状態変数 u を求めることができる．ここで，ある時刻 t における 2 つの GP2Y0A21YK0F のセンサ情報を $e_{1,t}, e_{2,t}$ ，それらに対応する状態変数を $u_{1,t}, u_{2,t}$ とおく．このとき，エージェントは，式 (6) のように，この状態変数 $u_{1,t}, u_{2,t}$ の列によって状態 s_t を次の通りに定義するものとする．

$$s_t = \{(u_{1,t}, u_{2,t}) | u_{1,t} = f_1 \circ f_2 \circ g'(e_{1,t}), u_{2,t} = f_1 \circ f_2 \circ g'(e_{e,t})\} \tag{19}$$

次に，エージェントに与える報酬について述べる．報酬は，タスクの進捗度に応じてエージェントに与えるものとする．このタスクでは，壁 A との距離 D_A がタスクの進捗度と相関があるものとする．ここで，ある時刻 t における壁 A との距離が $d_{A,t} \in D_A$ であるとする．この距離 $d_{A,t}$ に対応するセンサ情報，および状態変数は $e_{1,t} \in E_1$，$u_{1,t} \in U_1$ である．このとき，エージェントに与えられる報酬 r_t は，次式の通りに定義される．

$$\begin{aligned} r_t &= 20 - f_1 \circ f_2 \circ g'(e_{1,t}) \\ &(= 20 - u_{1,t}) \end{aligned} \tag{20}$$

両エージェントには，1 時刻毎に上式で決定される報酬 r_t を与えられる．特に，この報酬は毎時刻に与えられる即時報酬であることに注意されたい．これにより，提案手法を適用したエージェントは強化学習によって行動価値 $Q(s_t, a_t)$ を評価すると同時に，現在のタスクの進捗度を知ることに繋がる．

提案手法を適用したエージェントは，上式の報酬 r_t とそれぞれのセンサ情報 $e_{1,t}, e_{2,t}$ について 1 時刻毎に相関分析して，相関係数 $\rho_{1,t}, \rho_{2,t}$ を求めることになる．本実験では，提案手法を適用したエージェントがタスクを実行する上で必要であると判別する相関係数の閾値 ρ は，0.8 とする．

次に，エージェントの強化学習における行動について述べる．エージェントは全ての状態において，前方移動，右方移動，後方移動，左方移動，静止のいずれかの行動を選択可能である．これらの行動は，平行する 2 輪のモータの回転によって実現される．これらの行動のとき，モータは，床に完全に接地している時に 70 mm だけ移動する時間だけ回転する[11]．これは，現状態に隣接する状態に遷移できるだけの移動量に相当する行動である．ただし，壁に隣接する位置に相当する状態にエージェントが存在し，エージェントがそこで壁側の方向へ移動する行動を選択したときは，行動を実行しないものとする．すなわち，モータは回転動作をしないものとする．

エージェントの行動選択には $\epsilon-$greedy 法，行動評価には Q 学習を適用する．探査的な行動を選択する確率 ϵ，ステップサイズ・パラメータ α，割引率 γ は後記する表 5 の通りである．

以上のような実験設定のもと，エージェントにタスクを実行させる．エージェントは，壁 A および壁 B から最も遠い位置に相当する状態を初期状態として試行を開始する．このタスクの 1 試行の終了条件は，30000 回の行動選択とする．エージェントは，このような試行を 30 回だけ実行するものとする．

最後に，以上の実験設定を要約したものを表 5 に示す．

表 5: 実験設定の要約

項目	内容
1 試行の終了条件	1000 回の行動選択
総試行回数	30
選択可能な行動群 A	{ 前方移動，右方移動，後方移動，左方移動, 静止 }
遷移可能な状態数 $\|S\|$	11×11
行動選択	$\epsilon-$greedy
探査的な行動を選択する確率 ϵ	0.20
行動評価	Q 学習
ステップサイズ・パラメータ α	0.5
割引率 γ	0.5
行動価値 Q の初期値	0.0
相関係数の閾値 ρ	0.8

[11]強化学習における行動は，一般に状態認識のフィードバックに依らない．すなわち,「行動開始時の位置から 70 mm 移動するだけモータを回転させ続ける」というように行動を定義しない．強化学習における行動は，一般に動作的なものであることに注意されたい．

5.3.4 実験結果

本小節では，前小節で述べた設定をもとに実験した結果について述べる．本実験のタスクの終了条件はある一定の行動回数である．そのため，エージェントが 1 試行の間により多くの報酬値を獲得すれば良いといえる．そこで，本実験では，一般的な強化学習を適用したエージェントと提案手法を適用したエージェントが 1 試行の間にそれぞれ獲得した報酬値の累計を算出した．

はじめに，試行回数に対する両エージェントの累計報酬の推移を示すグラフを図 30 に示す．また，30 試行目の両エージェントの累計報酬の推移を示すグラフを図 31 に示す．図の通り，提案手法を適用したエージェントが獲得した累計報酬は，一般的な強化学習を適用したエージェントのものよりも，多くの試行において上回っていることが分かる．さらに，提案手法を適用したエージェントが獲得した累計報酬は安定しているのに対して，一般的な強化学習を適用したエージェントが獲得した累計報酬は不安定であることが分かる．これは，図 31 のように，試行の途中で一般的な強化学習を適用したエージェントの累計報酬の傾きが，しばしば下がることが起こるからである．したがって，提案手法を適用したエージェントの方が良い結果を示していると言える．

次に，試行回数に対するセンサ情報 $E_{1,t}, E_{2,t}$ と報酬値 R_t の相関係数 $\rho_{1,t}, \rho_{2,t}$ の推移を示すグラフを図 32 に示す．また，1 試行目，15 試行目，30 試行目の推移を示すグラフを，図 33．図 34 に示す図の通り，壁 A との距離に対応するセンサ情報 $e_{1,t} \in E_1$ と報酬値 $r_t \in R$ との相関係数 $\rho_{1,t}$ は，1 試行の最後には 0.9 付近に収束し始めていることが分かる．そして，その後のいずれの試行においても， 0.9 付近で収束していることが分かる．すなわち，その絶対値 $|\rho_{1,t}|$ が，1 試行目の開始直後から閾値 $\rho = 0.8$ を超え続けていることが分かる．また，壁 B との距離に対応するセンサ情報 $e_{2,t} \in E_2$ と報酬値 $r_t \in R$ との相関係数 $\rho_{2,t}$ は，1 試行目の開始直後に多少変動した後，-0.5 付近に収束し始めていることが分かる．そして，その後の試行では，-0.5 から正の方向に向かっている．すなわち，その絶対値 $|\rho_{2,t}|$ は，いずれの試行においても閾値を超えていないことが分かる．これは，提案手法を適用したエージェントが，1 試行の開始直後からその後の全ての試行にわたって，センサ情報 $e_{1,t} \in E_1$ に対応する状態変数 $u_{1,t} \in U_1$ のみで定義される状態にもとづいて行動を選択していたことを意味する．

次に，各試行において，提案手法を適用したエージェントが生成した限定利用に関する知識テーブルのレコードの数，および自律的判別に関する知識テーブルのレコード数の数位を示すグラフを図 35 に示す．図 35(a) の通り，限定的利用に関する知識テーブルのレコード数は，エージェントが遷移可能な状態数である 121 にしていることが分かる．限定的利用に関する知識テーブルのレコードは，前節で述べた通り，ある状態 s におけるある行動 a の価値 $Q(s, a)$ の更新回数 $N(s, a)$ を記録するものである．すなわち，これは，

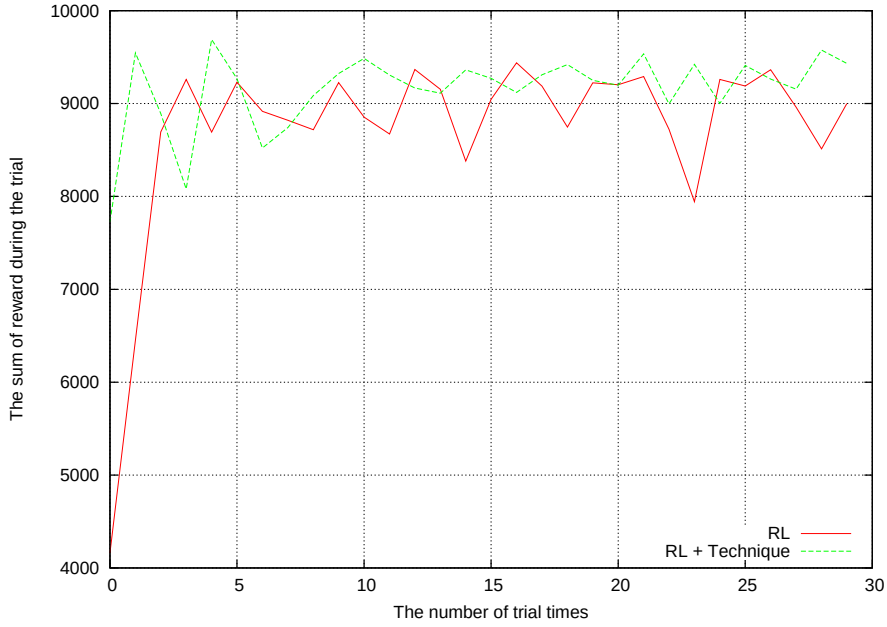


図 30: 試行回数に対する累計報酬の推移

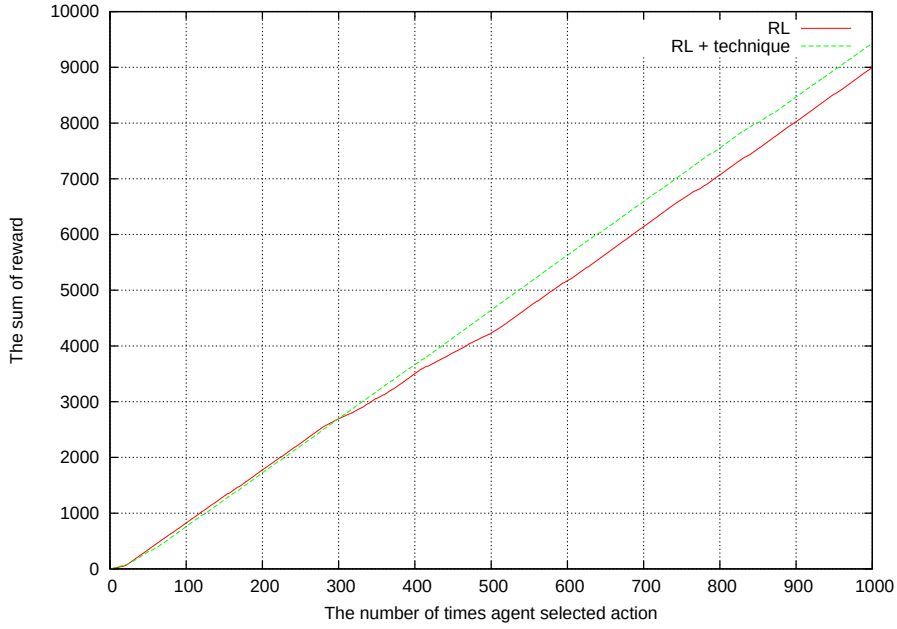


図 31: 30 試行目の両エージェントの累計報酬 $\sum_{t=1}^{t=1000} r_t$

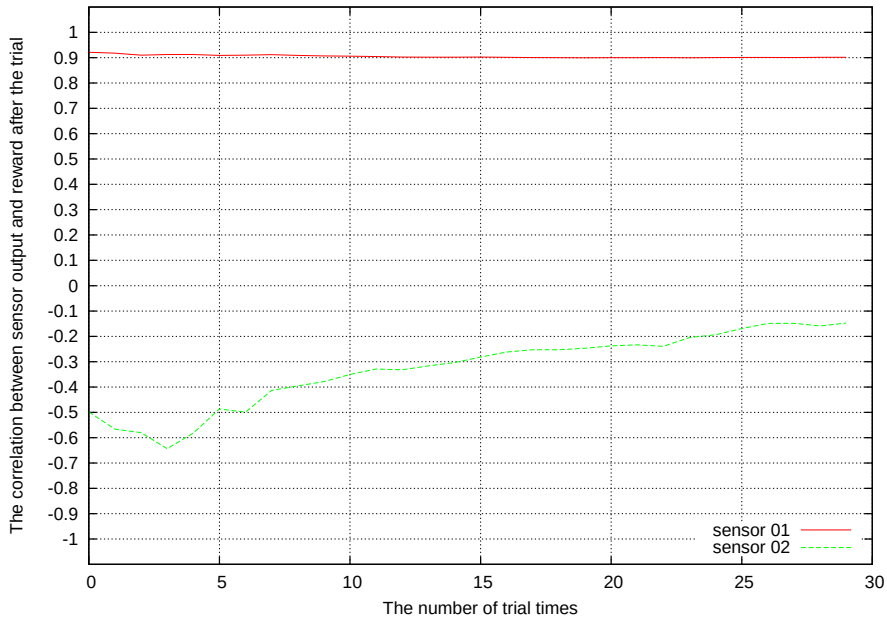


図 32: 試行回数に対する相関係数の推移（提案手法）

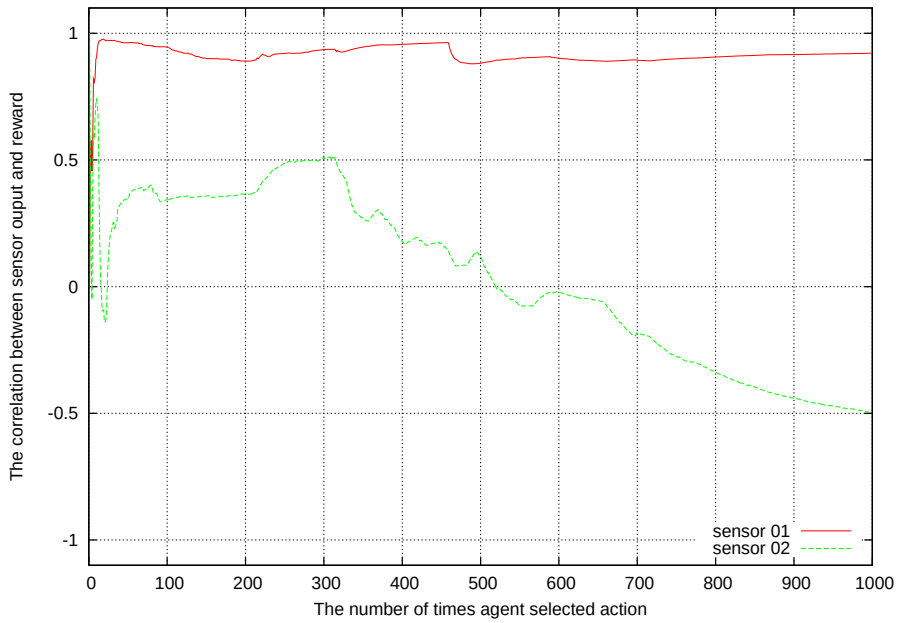
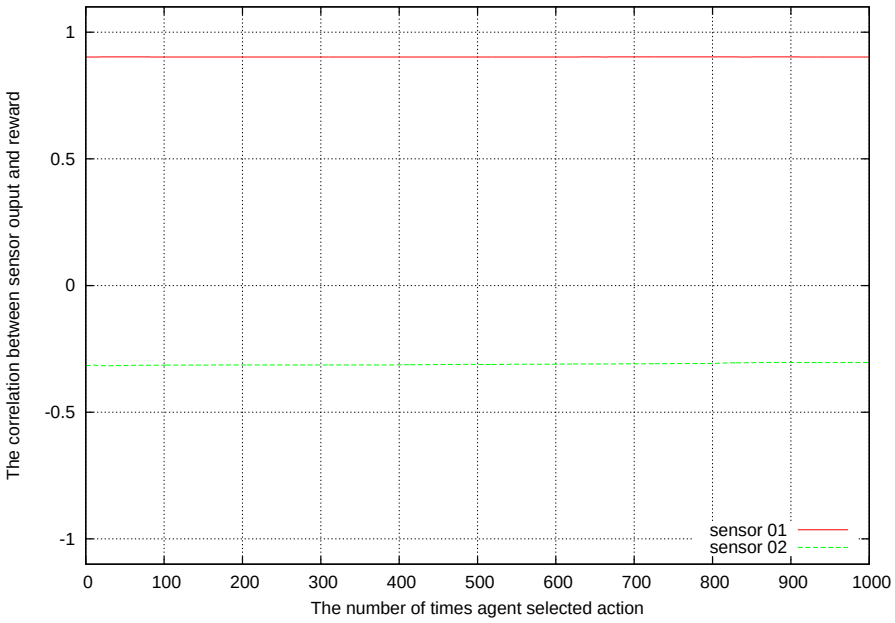
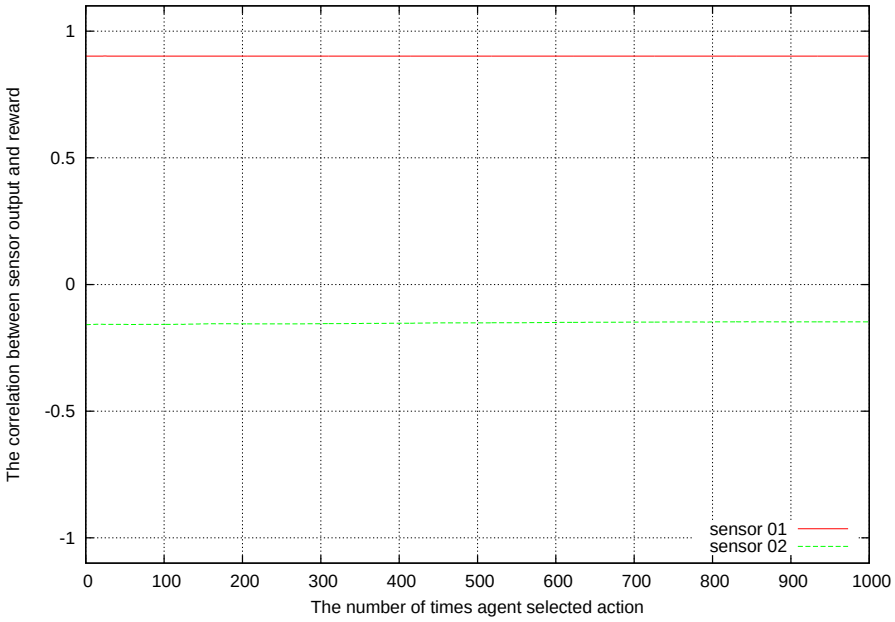


図 33: 1 試行目の相関係数 $\rho_{1,t}, \rho_{2,t}$ の推移（提案手法）



(a) 15 試行目



(b) 30 試行目

図 34: 15 試行目，30 試行目の相関係数 $\rho_{1,t}, \rho_{2,t}$ の推移

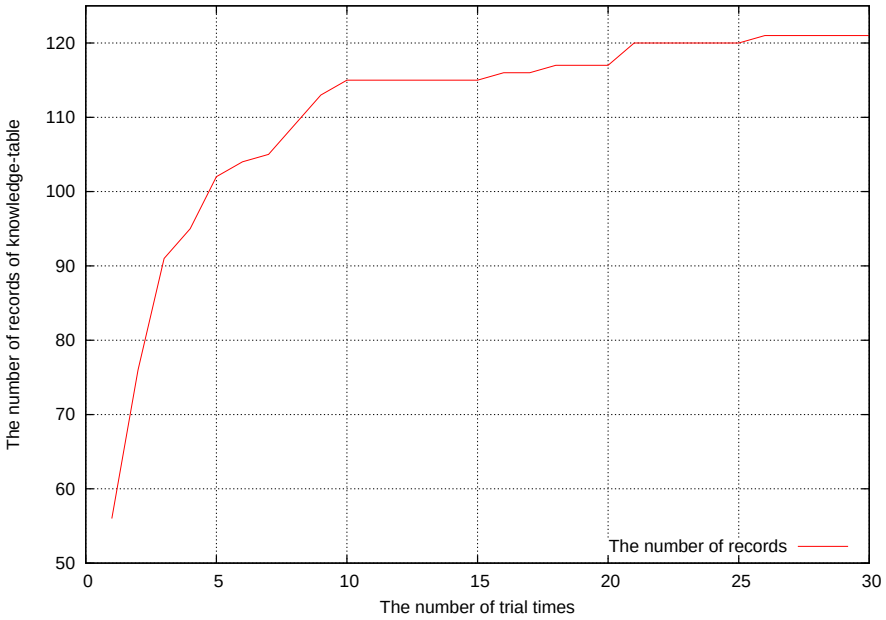
提案手法を適用したエージェントが 30 試行までに全ての状態に遷移したことを意味する．また，図 35(b) の通り，自律的判別に関する知識テーブルのレコード数は増え続けていることがわかる．しかしながら，図 32 に示した通り，相関係数 $\rho_{1,t}, \rho_{2,t}$ は収束している．このことから，それらの相関係数が十分に収束している時点で，それ以上のレコード数は必要がないと考えられる．すなわち，このレコード数は収束せずに増え続けていても問題はないと考えられる．

次に，5 試行目，10 試行目，15 試行目，20 試行目，25 試行目，30 試行目のときに，一般的な強化学習を適用したエージェントの壁 A および壁 B からの距離 $d_{A,t} \in D_A, d_{B,t} \in D_B$ の推移を示すグラフを図 36，図 37，図 38 に示す．また，提案手法を適用したエージェントのものを図 39，図 40，図 41 に示す．ただし，これらの距離は実際のものではなく，それらに対応するセンサ情報 $e_{1,t} \in E_1, e_{1,t} \in E_2$ から計算したものであることに注意されたい．そのため，ここではそれらの試行における両エージェントの $e_{1,t} \in E_1, e_{1,t} \in E_2$ の推移を示すグラフを図 42，図 43，図 44，および，図 45，図 46，図 47 に示す．ここで，これらのグラフの中で，値がインパルス状に大きく変化しているのは，センサ情報にノイズが混入したためである．ノイズの大きさは一定ではないが，特に，直前の距離に依らずに突然 80 cm を示している箇所は，ノイズが混入したものであると考えられる．

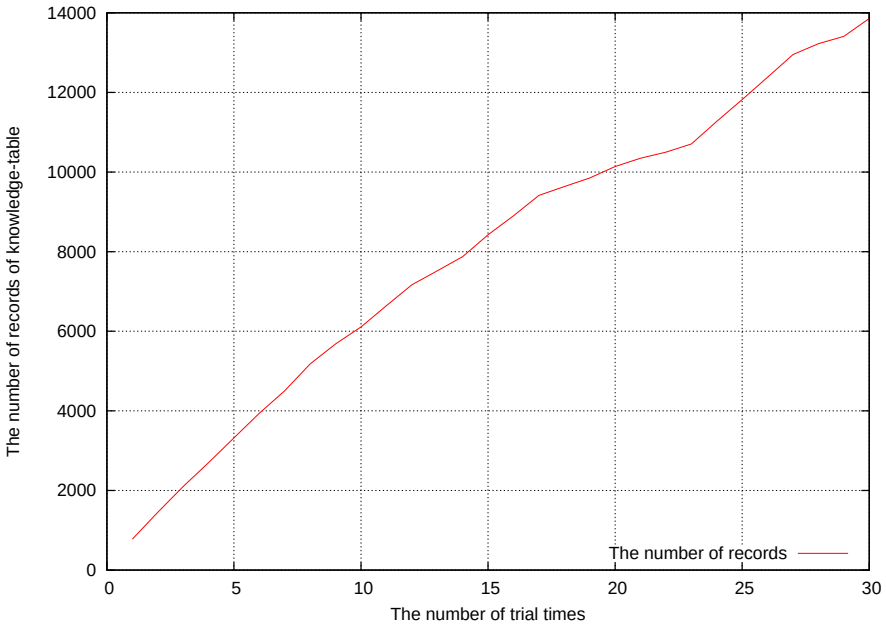
さて，図 36，図 37，図 38，および図 39，図 40，図 41 では，壁 A からの距離 d_A が 10 cm 付近であるとき，エージェントは目標状態に相当する壁 A の近傍に到達していることを意味する．これらの図の通り，いずれの試行においても，両エージェントは目標状態に到達できていることが分かる．しかしながら，一般的な強化学習を適用したエージェントは，幾つかの試行の間に，一時的に壁 A の近傍から離れていることがわかる．具体的に挙げれば，5 試行目においてはおよそ 300 回目から 400 回目の行動回数の区間，10 試行目においてはおよそ 150 目から 200 回目の行動回数の区間，15 試行目においては 450 回目から 550 回目および 700 回目から 750 回目の行動回数の区間，25 試行目においては 250 回目から 300 回目の行動回数の区間，30 試行目においては 300 回目から 350 回目および 400 回目から 500 回目の行動回数の区間である．一般的な強化学習を適用したエージェントは，これに伴って一時的に報酬を低く獲得している．これは，先に示した累計報酬の傾きの変化に相当する．これらのことは，センサ情報 $e_{1,t} \in E_1, e_{2,t} \in E_2$ についても同様である．

5.3.5 考察

これらの実験結果から，第 1 節で述べたこととほとんど同様の考察が言えると考えられる．すなわち，一般的な強化学習を適用したエージェントの場合，目標状態に一度到達してその近傍の状態で学習するが，その途中で未探

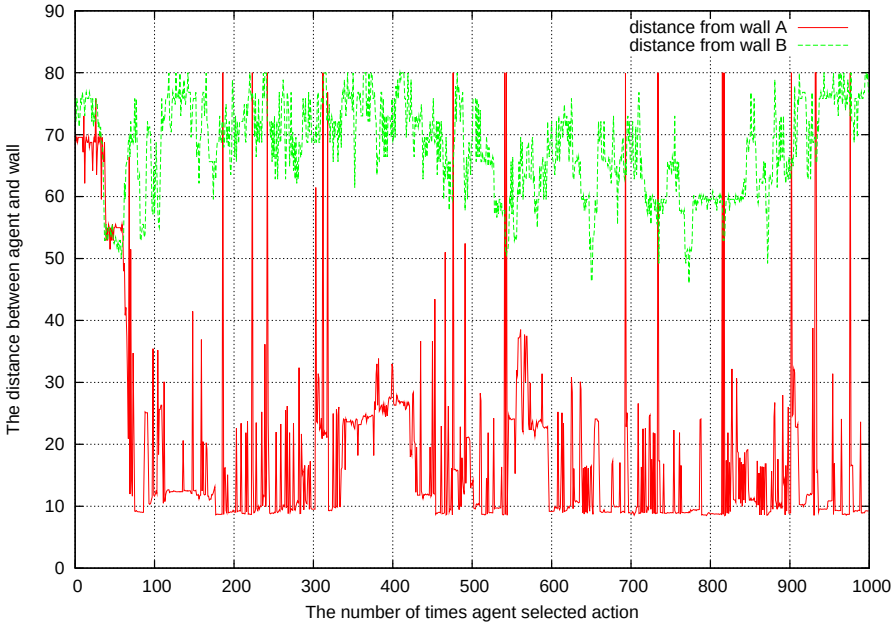


(a) 限定的利用に関する知識テーブルのレコード数

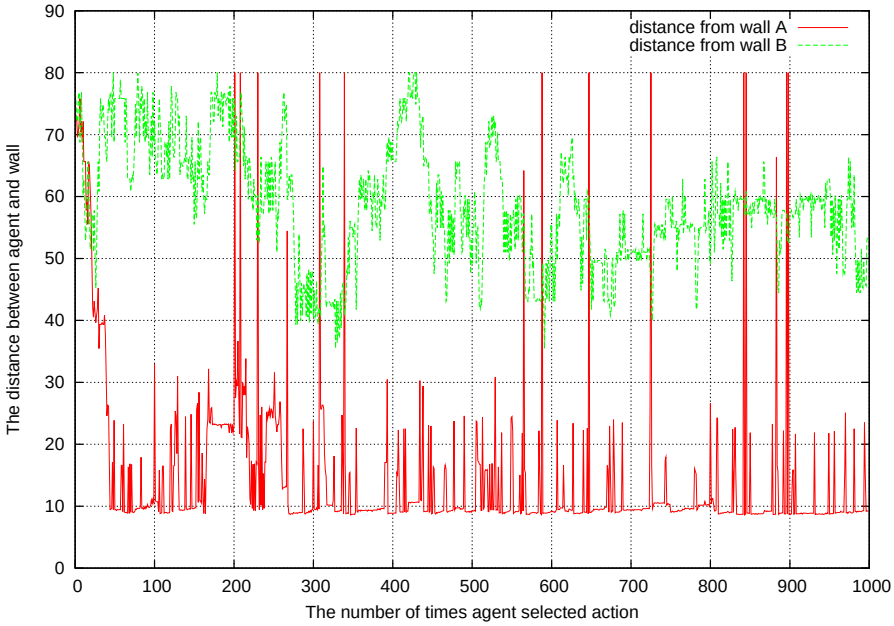


(b) 自律的判別に関する知識テーブルのレコード数

図 35: 30 試行目までに提案手法を適用したエージェントが獲得した知識

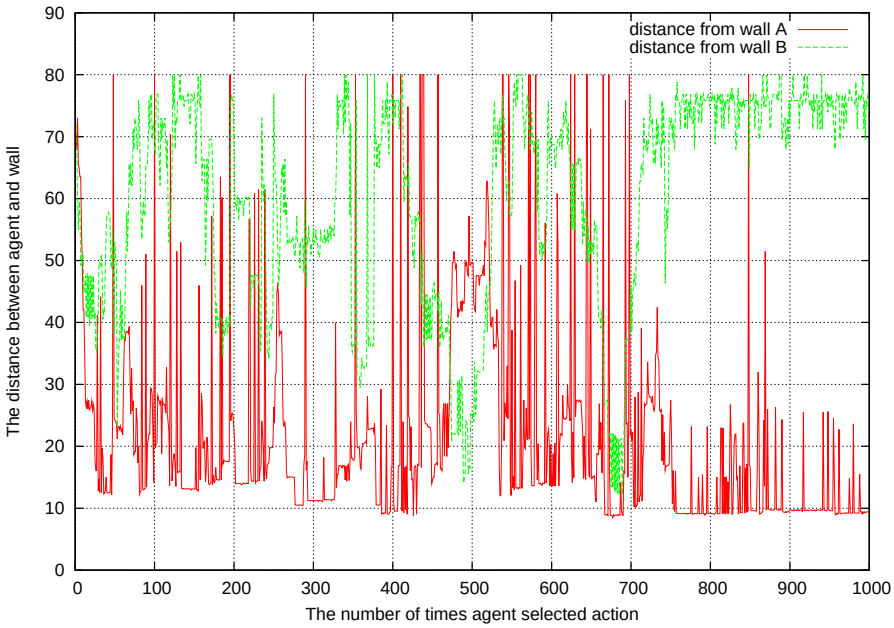


(a) 5 試行目

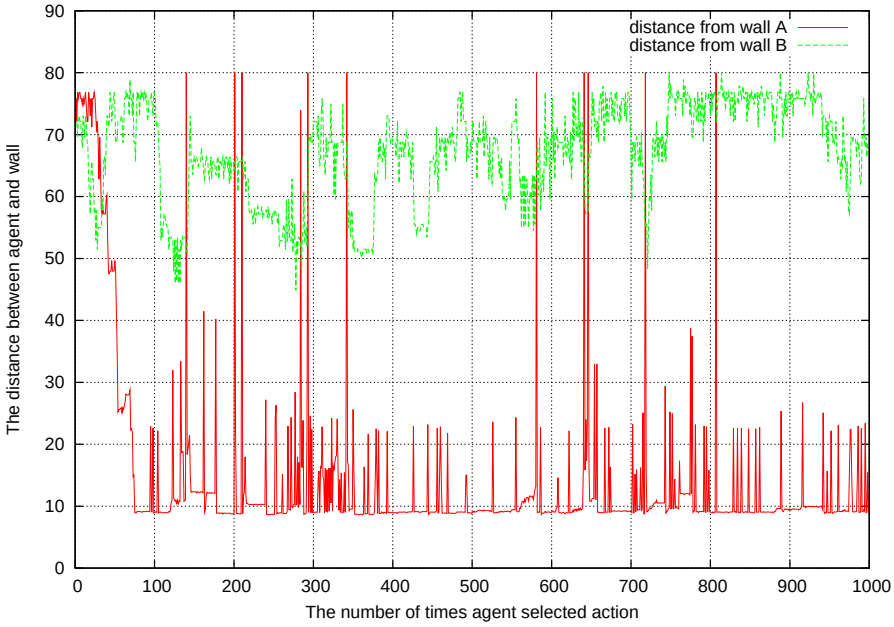


(b) 10 試行目

図 36: 一般的な強化学習を適用したエージェントの壁からの距離の推移

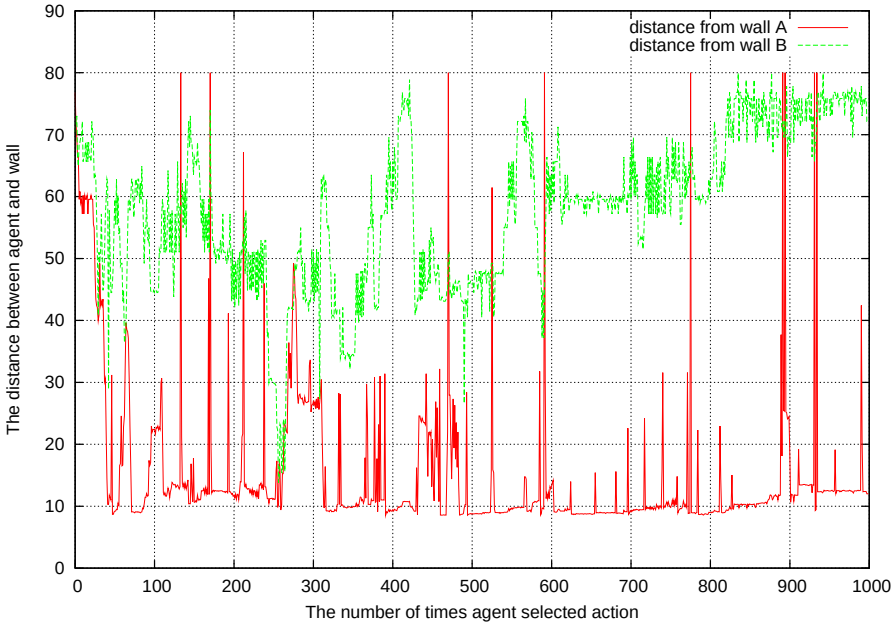


(c) 15 試行目

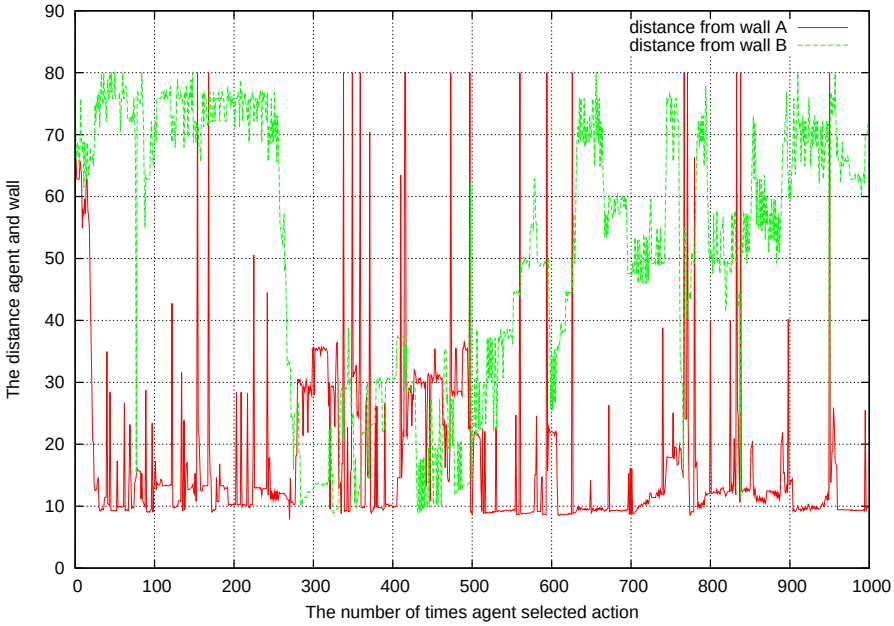


(d) 20 試行目

図 37: 一般的な強化学習を適用したエージェントの壁からの距離の推移

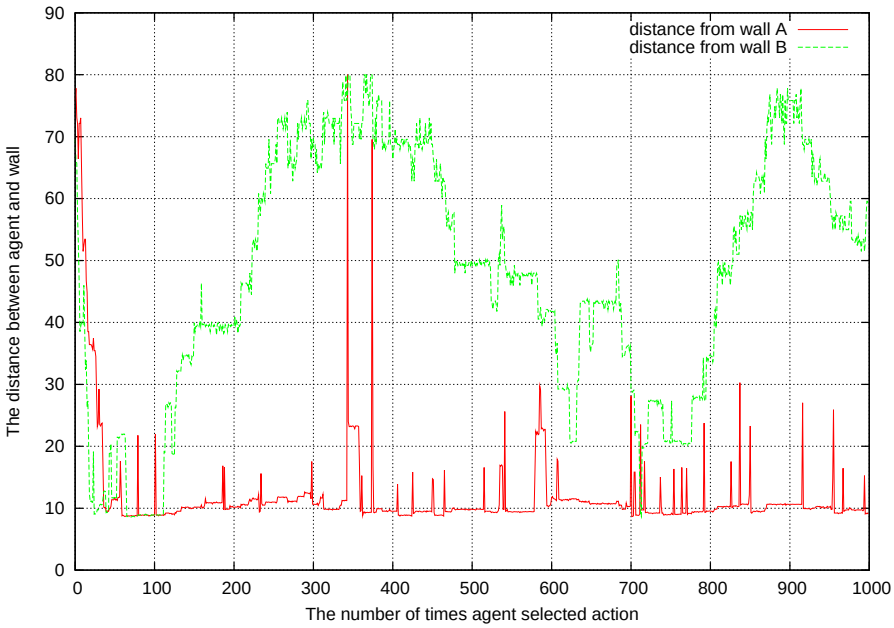


(e) 25 試行目

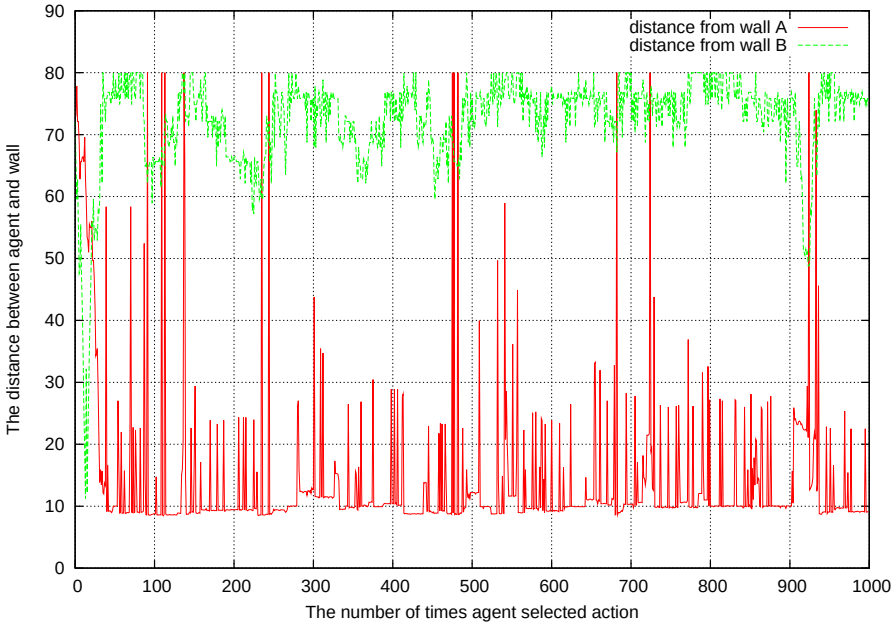


(f) 30 試行目

図 38: 一般的な強化学習を適用したエージェントの壁からの距離の推移

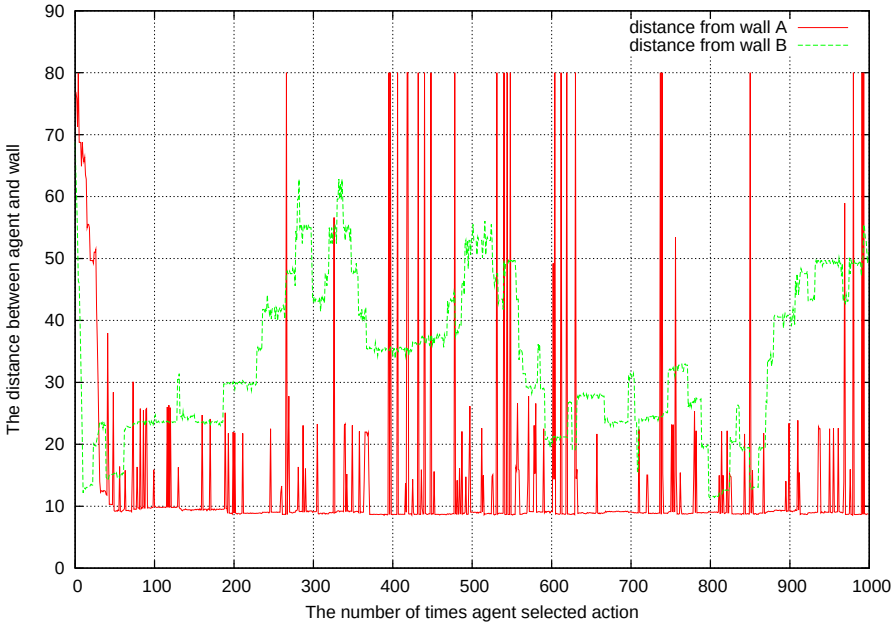


(a) 5 試行目

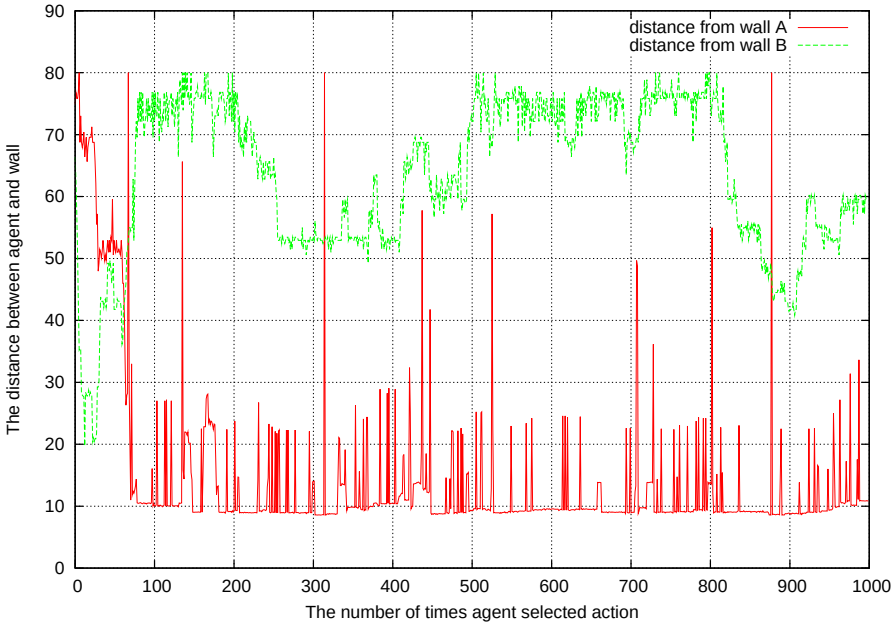


(b) 10 試行目

図 39: 提案手法を適用したエージェントの壁からの距離の推移

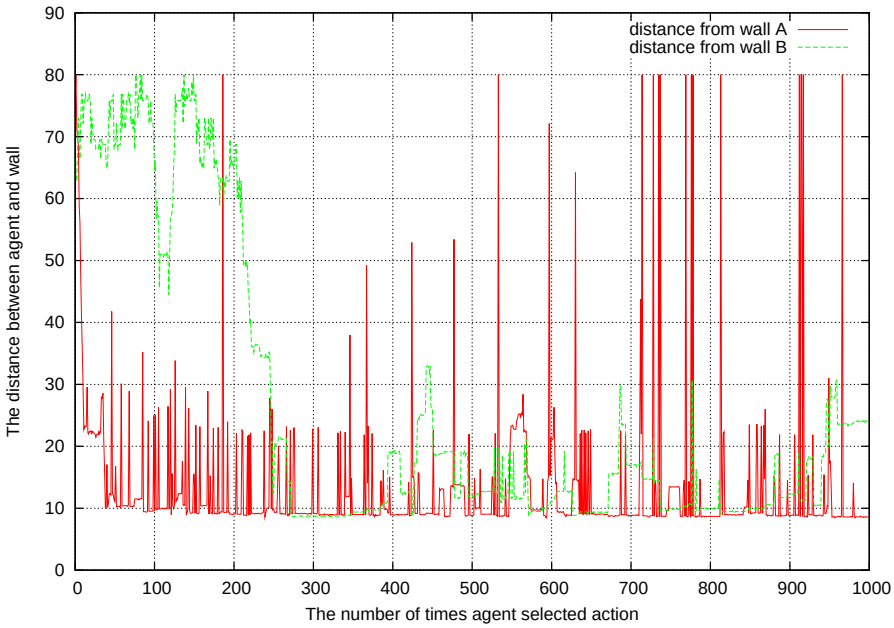


(c) 15 試行目

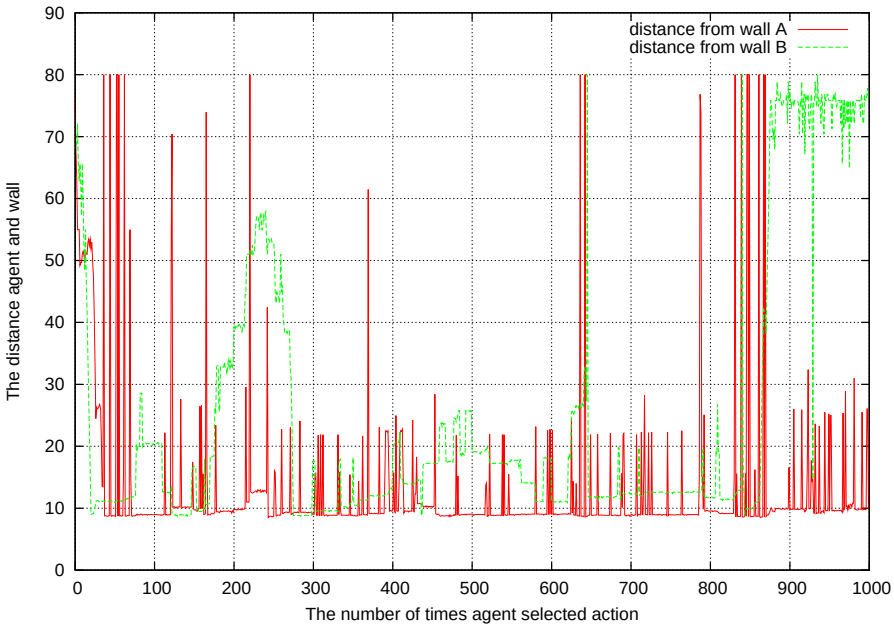


(d) 20 試行目

図 40: 提案手法を適用したエージェントの壁からの距離の推移

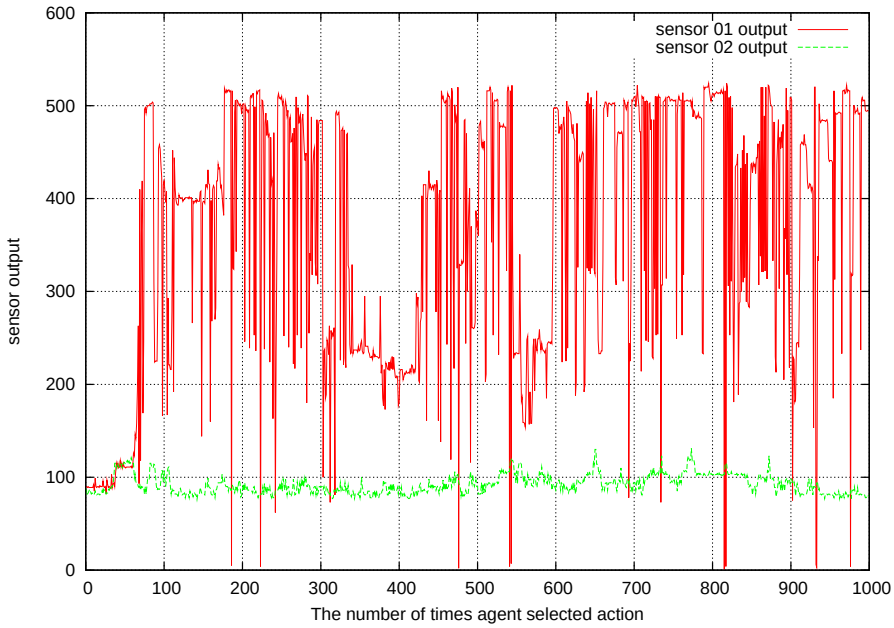


(e) 25 試行目

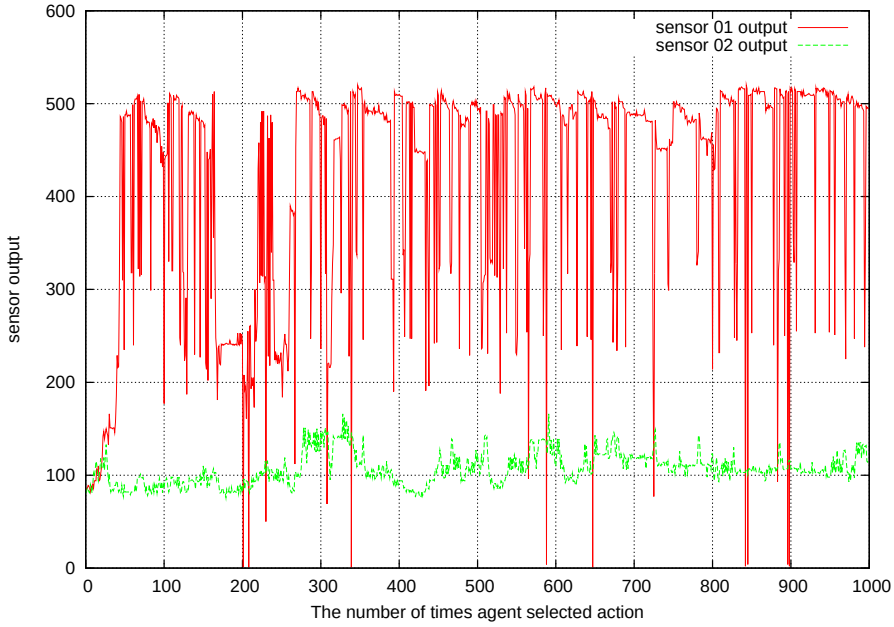


(f) 30 試行目

図 41: 提案手法を適用したエージェントの壁からの距離の推移

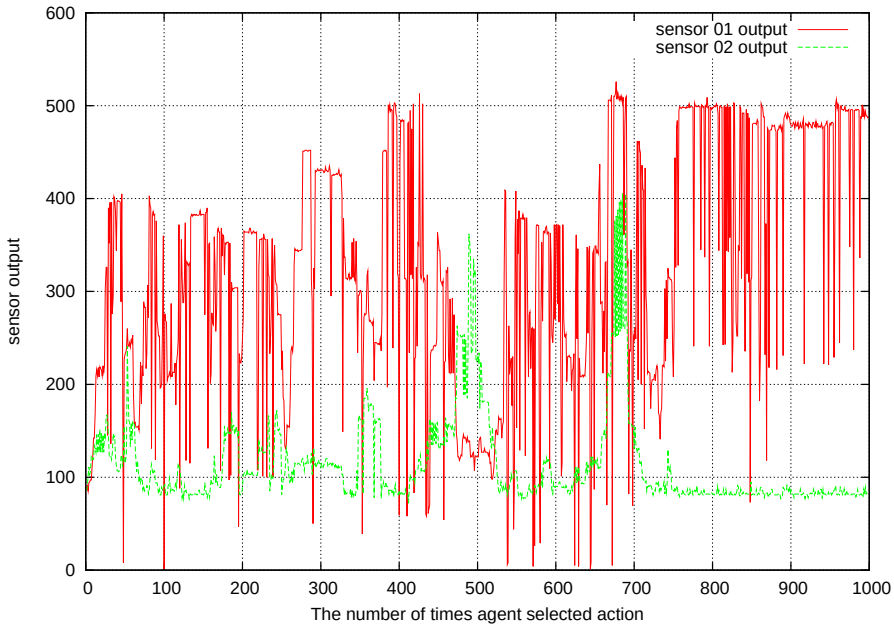


(a) 5 試行目

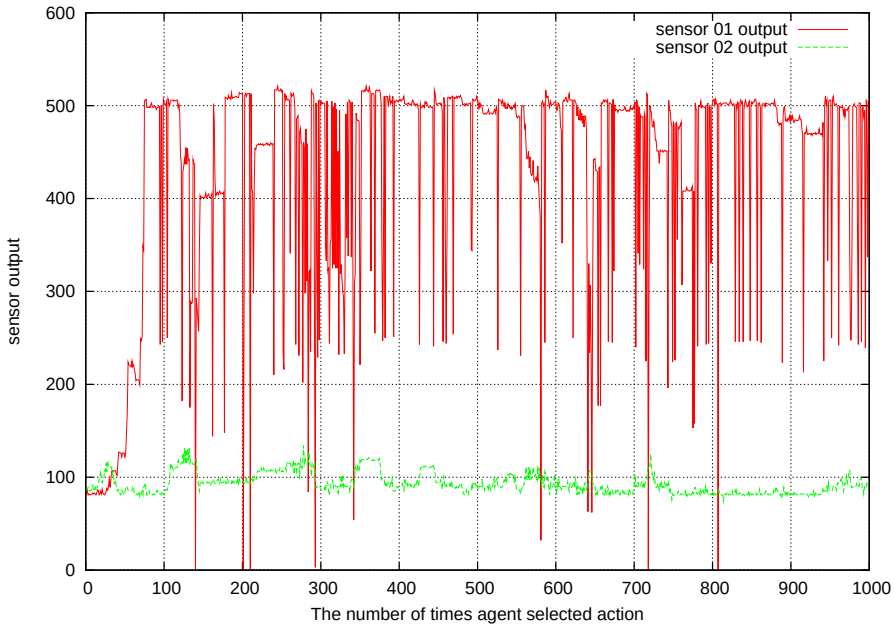


(b) 10 試行目

図 42: 一般的な強化学習を適用したエージェントのセンサ情報の推移

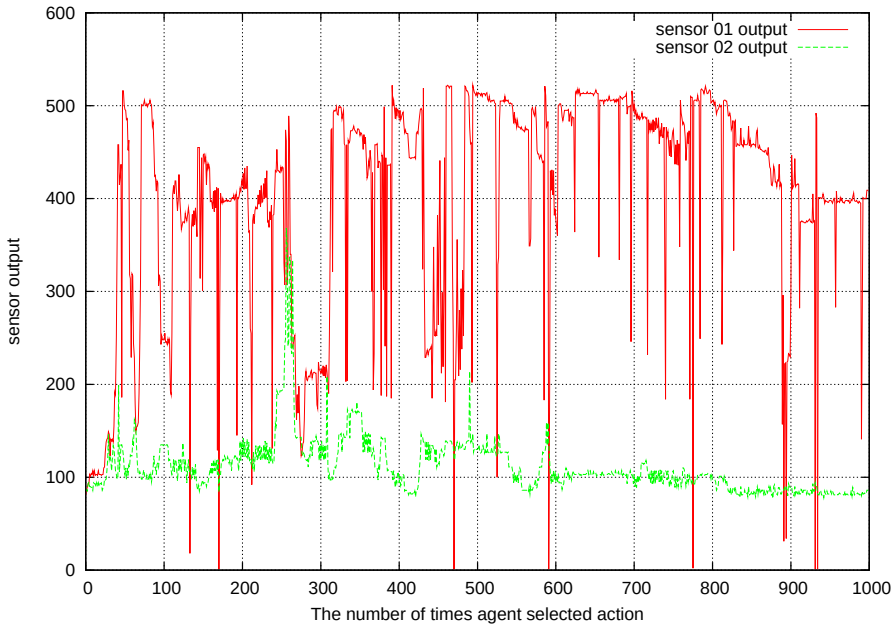


(c) 15 試行目

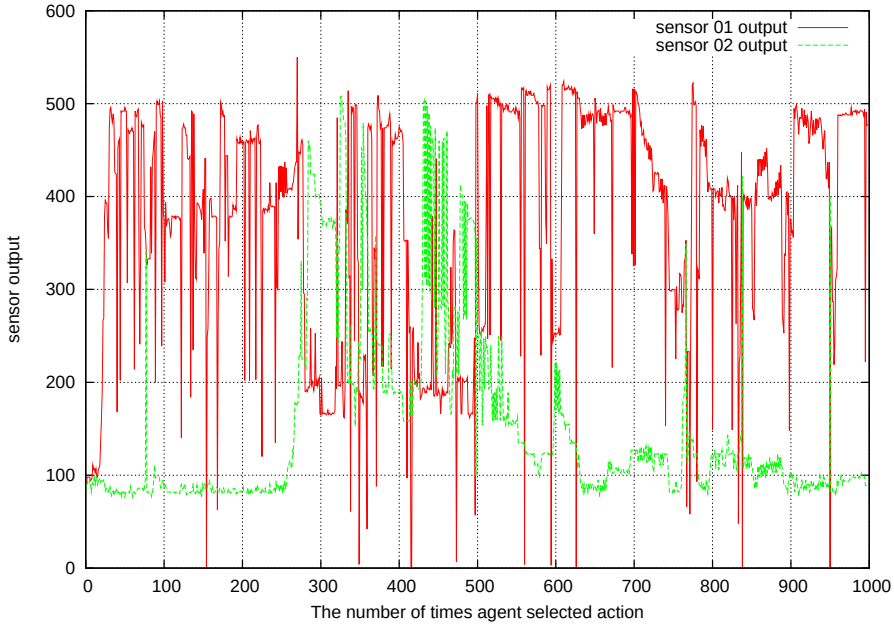


(d) 20 試行目

図 43: 一般的な強化学習を適用したエージェントのセンサ情報の推移

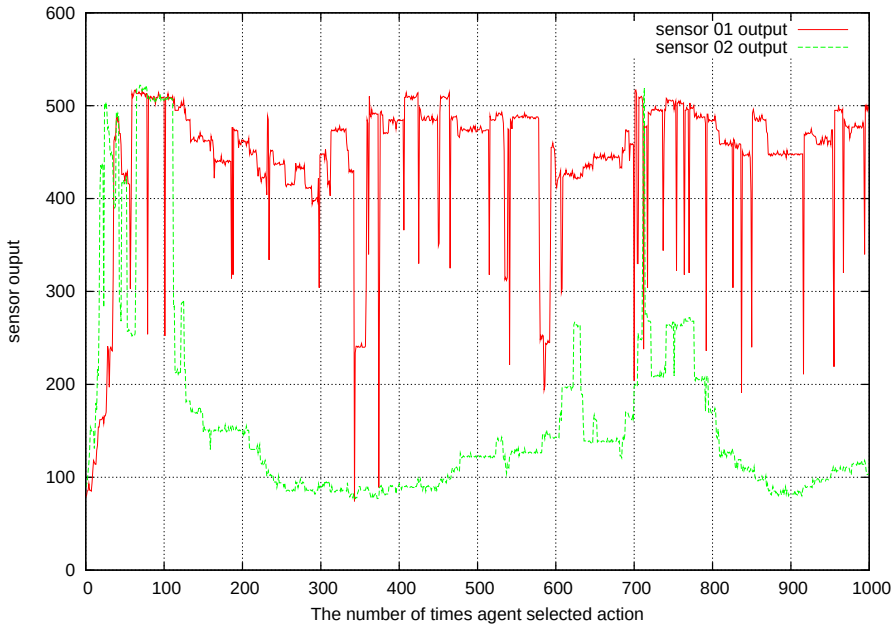


(e) 25 試行目

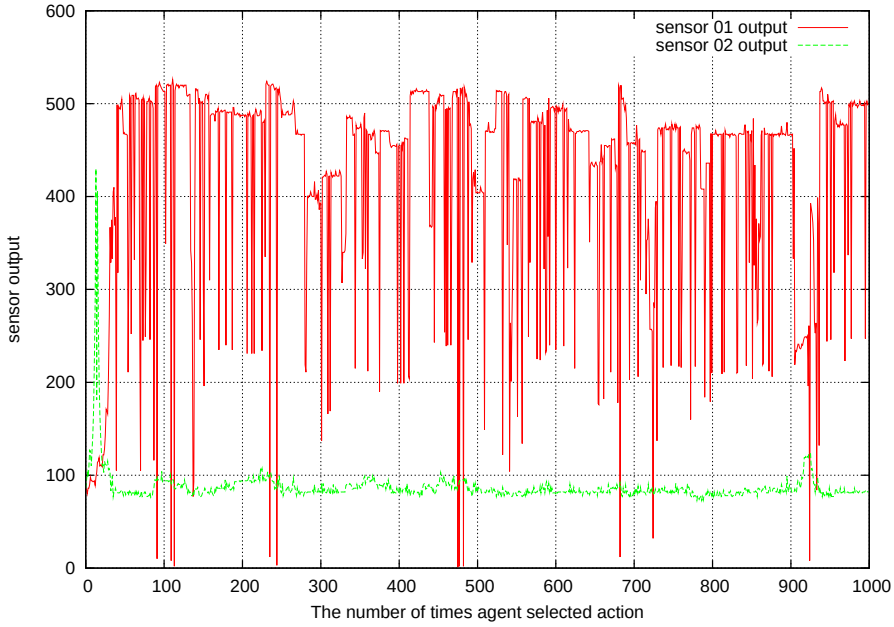


(f) 30 試行目

図 44: 一般的な強化学習を適用したエージェントのセンサ情報の推移

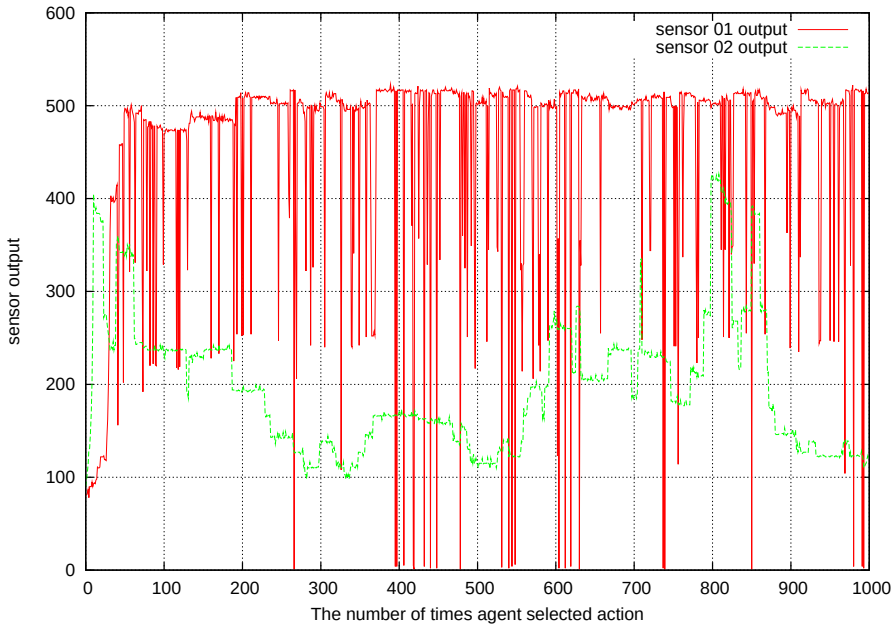


(a) 5 試行目

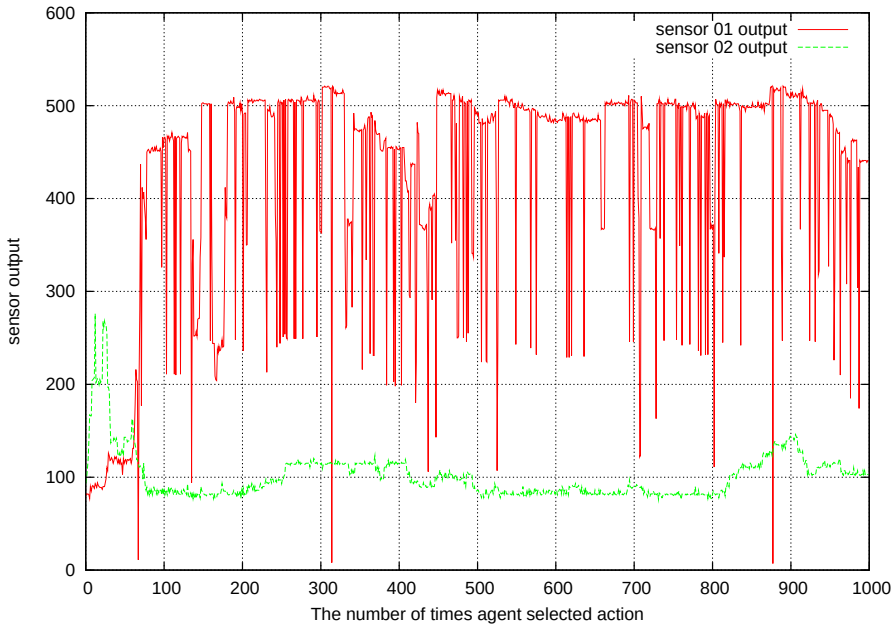


(b) 10 試行目

図 45: 提案手法を適用したエージェントのセンサ情報の推移

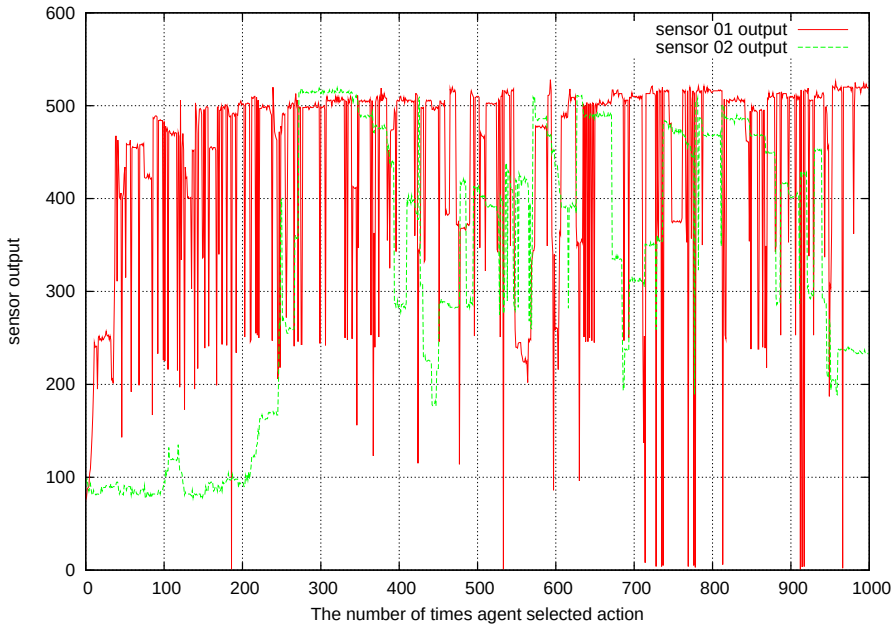


(c) 15 試行目

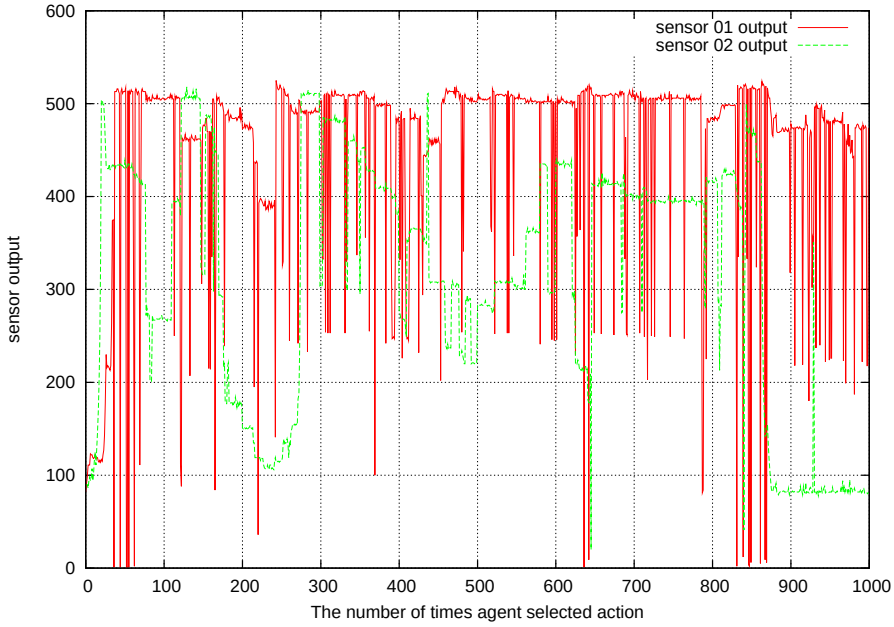


(d) 20 試行目

図 46: 提案手法を適用したエージェントのセンサ情報の推移



(e) 25 試行目



(f) 30 試行目

図 47: 提案手法を適用したエージェントのセンサ情報の推移

索な状態群へ迷い込んだことが考えられる．このことは，特に 15 試行目と 24 試行目における累計報酬の大きな下がり方として現れている．また，それらの試行における壁 A からの距離の推移からも同様に言える．

さて，このことの他に，今回の実験ではセンサ情報に混入したノイズによっても，両エージェントの結果の違いが現れていると考えられる．前小節で述べた通り，両エージェントのセンサ情報や距離が，しばしばインパルス状に大きく変化していることが分かる．このとき，両エージェントは本来の状態 s_t ではなく，それとは別の何らかの状態 s'_t を誤って認識をしている．この誤認識による両エージェントの行動選択と行動評価について考える．

はじめに，一般的な強化学習を適用したエージェントについて考える．一般的な強化学習を適用したエージェントは，ある状態 s_t における任意の行動 a お価値 $Q(s_t, a)$ を参照して行動選択する．このとき，そのエージェントは本来 $Q(s_t, a)$ を参照すべきところを，$Q(s'_t, a)$ を参照することになる．これにより，エージェントは状態の誤認識と同時に，誤った行動選択をすることになることが考えられる．また，エージェントの行動評価の方法が Q 学習であることを考える．Q 学習では，式 (2) の通り，次の時刻の状態における行動価値をフィードバックしている．つまり，ある時刻 t においてエージェントが次の状態 s_{t+1} を s'_{t+1} であると誤認識したとき，その時刻 t における行動選択の後では $\max_{a \in A(s_t)} Q(s'_{t+1}, a)$ を不当にフィードバックしてしまう．すなわち，$Q(s_t, a_t)$ を不当に評価してしまうことが考えられる．さらに，その次の時刻 $t+1$ では，その時刻 $t+1$ における行動選択の後では $Q(s'_{t+1}, a_{t+1})$ を不当に評価してしまうことが考えられる．これにより，一般的な強化学習を適用したエージェントが状態 s_t, s'_{t+1} において適切な行動を選択をするためには，誤認識によって不当な評価をしてしまった価値 $Q(s_t, a_t), Q(s'_{t+1}, a_{t+1})$ を再度学習しなければならないことが考えられる．

次に，提案手法を適用したエージェントについて考える．提案手法を適用したエージェントの場合も同様に，状態を誤認識してしまった時には，価値 $Q(s_t, a_t), Q(s'_{t+1}, a_{t+1})$ を不当に評価してしまう．しかしながら，提案手法を適用したエージェントは，提案手法が機能しているとき，式 (11) に従って，再定義された状態 s^* における任意の行動 a 価値 $Q^*(s_t, a)$ を参照して行動選択する．すなわち，その時刻 t における同じ状態変数 $u_{1,t}$ をもつ他の状態の行動価値を加重平均してから参照するため，不当な価値の評価による影響を軽減することができる可能性がある．これにより，提案手法を適用したエージェントは，誤認識によって不当な評価をしてしまった価値を再度学習せずとも，適切な行動選択ができることが考えられる．

したがって，これらのことから，提案手法を適用したエージェントが獲得した報酬の累計が，一般の強化学習を適用したエージェントのものよりも上回ることになったのである．

6　まとめ

本研究では，はじめに，学習を適用したロボットがセンサ情報を自律的に選択する必要があることを述べた．そして，本研究では，タスクの進捗度と環境の物理量との間に相関があることに着目し，これによって必要なセンサ情報を判別することをアプローチとして述べた．さらに，そのアプローチに基づき，強化学習の枠組みの中でその具体的な選択方法と利用方法について考え，それらを取り入れた学習機構を提案した．それに続く 3 つの実験では，一般的な強化学習を適用したエージェントと提案手法を適用したエージェントを比較し，いずれも後者が前者よりも良い結果であったことを示した．特に，実際のロボットを使用した実験では，センサ情報のノイズが混入する場合に対しても，有効に機能する可能性があることを示した．これらのことから，本研究の提案手法の有効性を示すことができたと考えられる．

今後の課題としては，他のセンサ情報の自律的な選択方法の探求，自律的判別に関する知識テーブルのレコード数の削減，他の環境における提案手法の有効性の検証が挙げられる．

はじめに，他のセンサ情報の自律的な選択方法について述べる．本研究では，ピアソンの積率相関係数を求める相関分析を利用している．この相関分析は，ある 2 変量の関係がどれだけ線形関係に近しいかを調べるものである．これ以外の方法に，重回帰分析が利用できると考えられる．重回帰分析は，多変量解析の方法の 1 つであり，幾つかの説明変数とある 1 つの目的変数の 1 次式の関係を調べるものである．すなわち，説明変数にセンサ情報，目的変数に報酬値として，それらを関係を分析することが考えられる．特に，相関分析が因果関係を言及できないのに対し，重回帰分析は因果関係を言及できる．このことから，センサ情報の必要性をより厳密に分析することができると考えられる．

次に，自律的判別に関する知識テーブルのレコード数の削減について述べる．本研究では，レコードの項の組み合わせが重複しない限り，際限なくレコードを追加するものとしている．これは，ロボットのメモリが有限であることを考えると，実用的ではないと言える．したがって，今後は何らかの基準によって追加を抑制するか，削除することを考える必要がある．これについては，一般に良く知られる LFU(Least Frequently Used)，LRU(Least Recently Used) アルゴリズムのような基準が考えられる．

最後に，他の環境における提案手法の有効性の検証について述べる．本研究では，いずれの実験も格子状の状態遷移図で記述できるような環境で検証した．すなわち，本研究の提案手法の有効性が，そのような環境でしか保証できないと可能性があると十分に考えられる．したがって，今後は別の状態遷移の構造をもつ環境でさらに検証し，どのような環境においても本研究の提案手法の有効性が広く保証できることを目指す必要があると考えられる．

A　付録：実験機ロボットの図面

本章では，付録として第 5 章 2 節の実験で使用したロボットの投影図，組立図および回路図を記載する．投影図，組立図の縮尺は，可能な限り拡大して見易くするため，それぞれ異なっていることに注意されたい．ロボットの詳細な寸法や電子回路の構成は，本章で参照されたい[12] ．

A.1　投影図

本研究の実験で使用したロボットを構成する各部品の投影図を記載する．ここで，以下に載せる投影図は，第三角法に従って製図したものである．また，正面図の上下もしくは左右が対称である場合は，慣例に従って，右側面図や上面図の記載は割愛する．

はじめに，ロボットの第 1 階層の投影図を図 48 に示す．第 1 階層は，ロボットにセンサを搭載するための階層である．この階層には，距離を計測する赤外線センサの GP2Y0A21YK0F が搭載されている．図中の想像線（二点鎖線）で描かれた物体が，その赤外線センサ GP2Y0A21YK0F である．上図では，GP2Y0A21YK0F の概形とその寸法のみしか記載していない．GP2Y0A21YK0F の詳細な図面は，後記した文献を参考されたい [26].

次に，ロボットの第 2 階層の投影図を図 49 に示す．ただし，上面図および右側面図は，回路や電子素子などの点数が多いために非常に複雑となるので，記載は割愛する．第 2 階層は，ロボットに電子回路を搭載するための階層である．この階層には，Armadillo-300，Arduino UNO，AGB65-RSC2 および AGB65-232C が搭載されている．図中の Armadillo-300 は，回路を配置した方向が分かるように，UBS のコネクタ，LAN ケーブルのコネクタ，コンパクト・ディスクのコネクタの概形を右から順に描いた．Armadillo-300，Arduino UNO，AGB65-RSC2 および AGB65-232C の詳細な図面は，後記した文献を参考されたい [30] [31]．第 2 階層に配置した回路は，それぞれ電気的に繋がっている．これらの回路の構成については，後節の回路図を参考にされたい．

最後に，ロボットの第 3 階層の投影図を以下に示す．第 3 階層は，ロボットの移動機構のための階層である．図 50 の通り，第 3 階層には H ブリッジ回路を搭載している．H ブリッジ回路は，DC モータの回転方向を制御するものである．これについては，後節の回路図を参考にされたい．また，第 3 階層の裏面には，オムニホイール・モータを搭載している．オムニホイール・モータは，能動回転方向に対して垂直な方向に受動回転できる．したがって，図の通りに十字型に搭載しても，ロボットは平行する 2 輪を同じ方向に回転させることで，問題なく移動することができる．この搭載位置の寸法は，後節の組立図を参考にされたい．

[12] 本付録に記載した図面は,『初心者のための機械製図（第 2 版）』を参考にして製図した [24] .

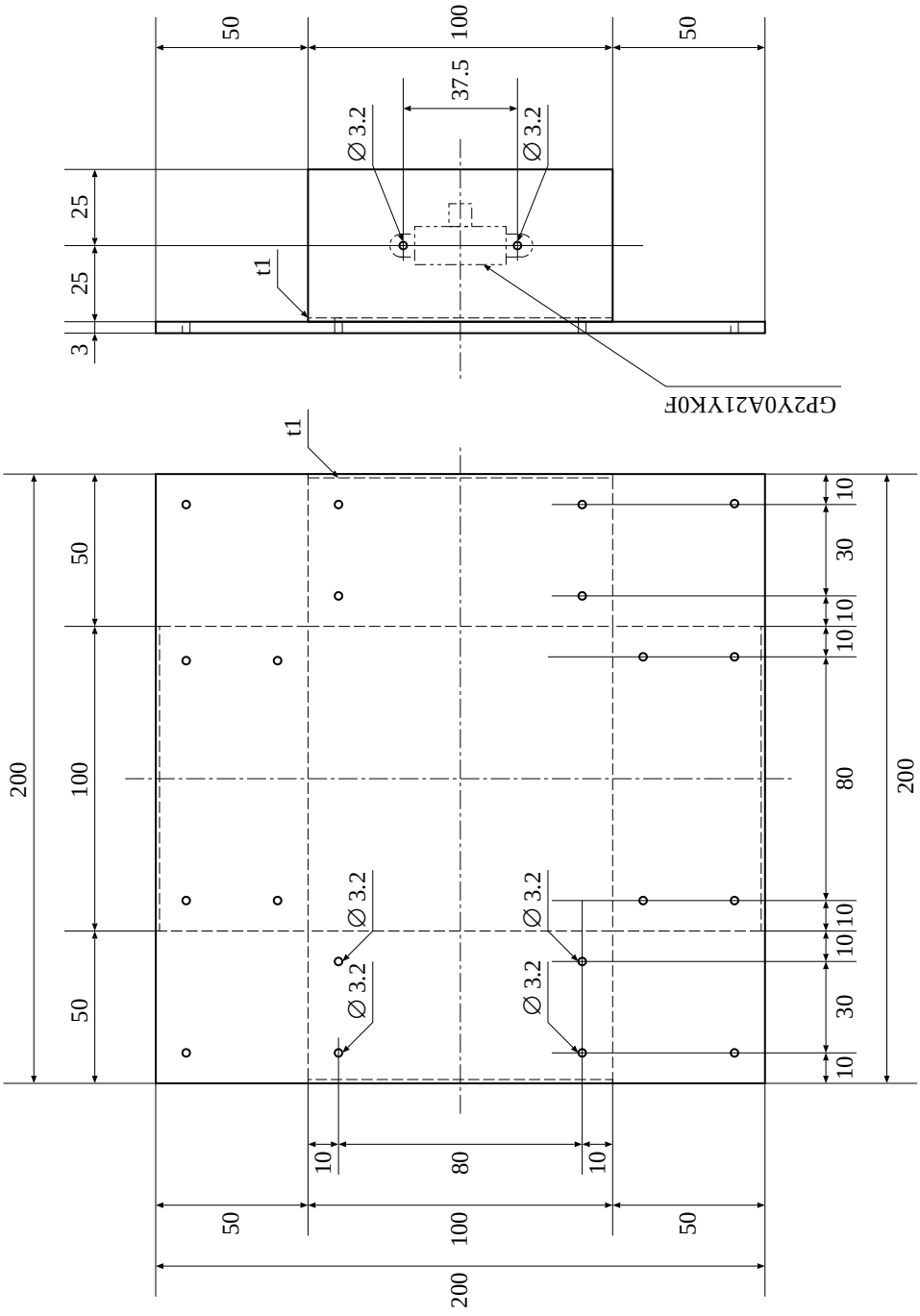


図 48: 第 1 階層（最上階層）の投影図（反時計回りに 90 度回転）

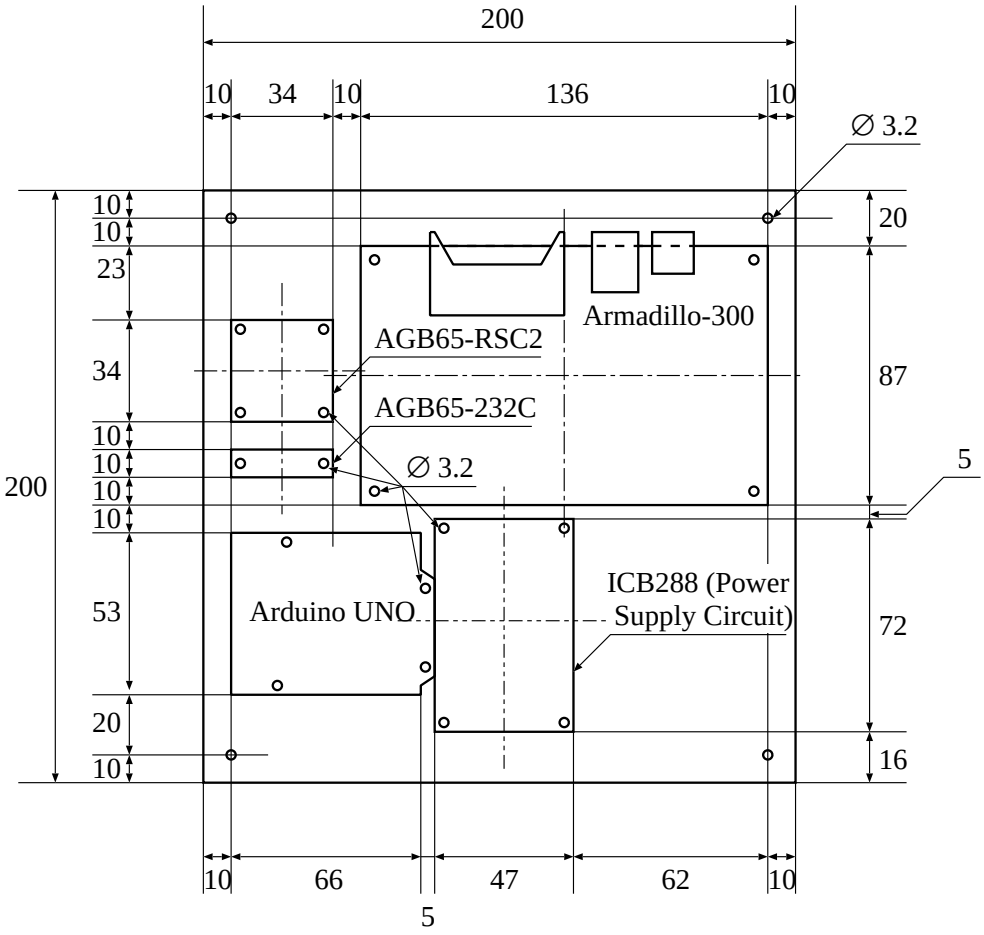


図 49: 第 2 階層（中間層）の投影図

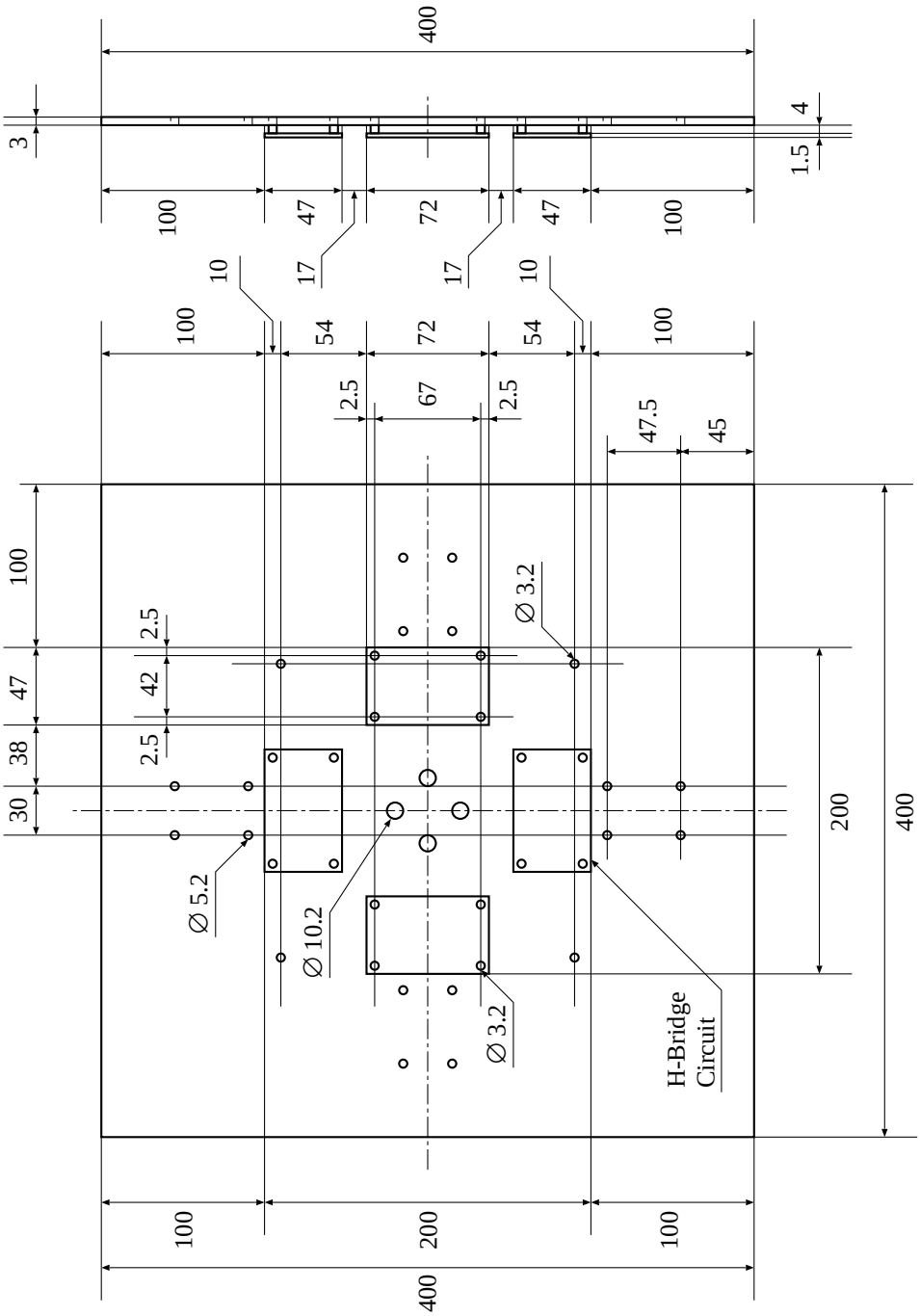


図 50: 第 3 階層（最下階層）の投影図（反時計回りに 90 度回転）

A.2　組立図

本研究の実験で使用したロボットの組立図を記載する．ここで，以下に載せる組立図は，第三角法に従って製図したものである．

はじめに，ロボットの上面図を図 51 に示す．図の通り，ロボットは幅 400 (mm)，奥行き 400 (mm) の広さをもつ．第 1 階層，第 2 階層，第 3 階層は，ロボットの機体の中央で階層的に組み立てられる．かくれ線（破線）で描かれた物体は，オムニホイール Urethane-Omni TYPE2581 と DC モータ RP380-ST である[13]．これは，第 3 階層の裏側で十字方向に取り付けられている．中央に存在する 4 つの穴は，それぞれのモータの電力線を通すためのものである．

次に，ロボットの正面図を図 52 に示す．図の通り，ロボット 262 (mm) の高さを持つ．また，赤外線センサ GP2Y0A21YK0F は 234 (mm) の高さでロボットに搭載されている．さらに，赤外線センサ GP2Y0A21YK0F は第 3 階層の端から中心側へ 100 mm のところに搭載されている．この理由は，第 5.3 節で述べた通りである．

[13]これは，株式会社の土佐電子が販売している商品である．オムニホイールは，同社が販売している「オムニホイール TD-80」，モータは，株式会社タミヤが販売している 380 シリーズのギヤード・モータである．図 51，図 52 では，それぞれの概形とそれらの特徴的な部分の寸法のみしか記載していない．これらの詳細な図面は，後記した文献を参考されたい [28] [29].

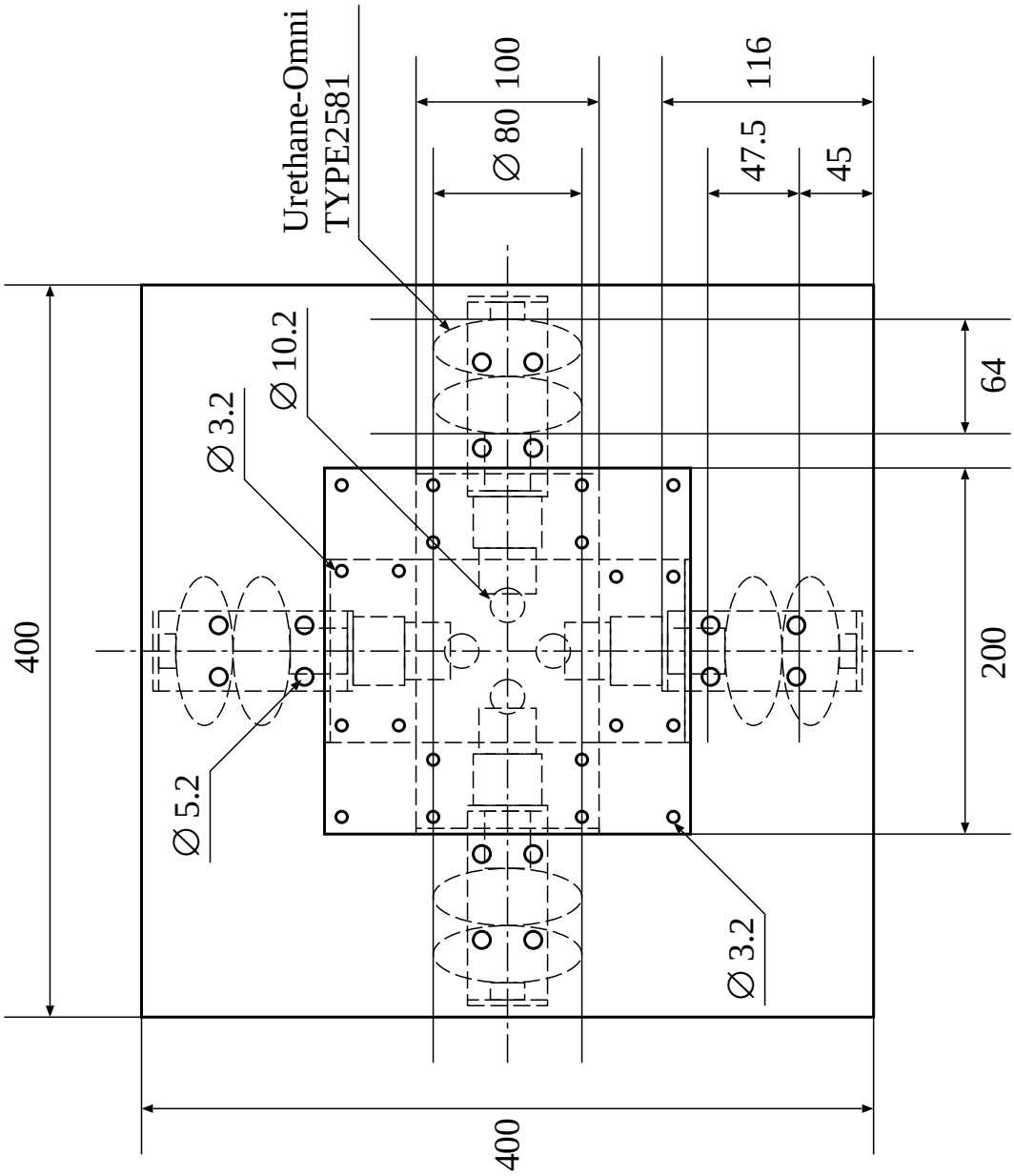


図 51: ロボットの上面図（反時計回りに 90 度回転）

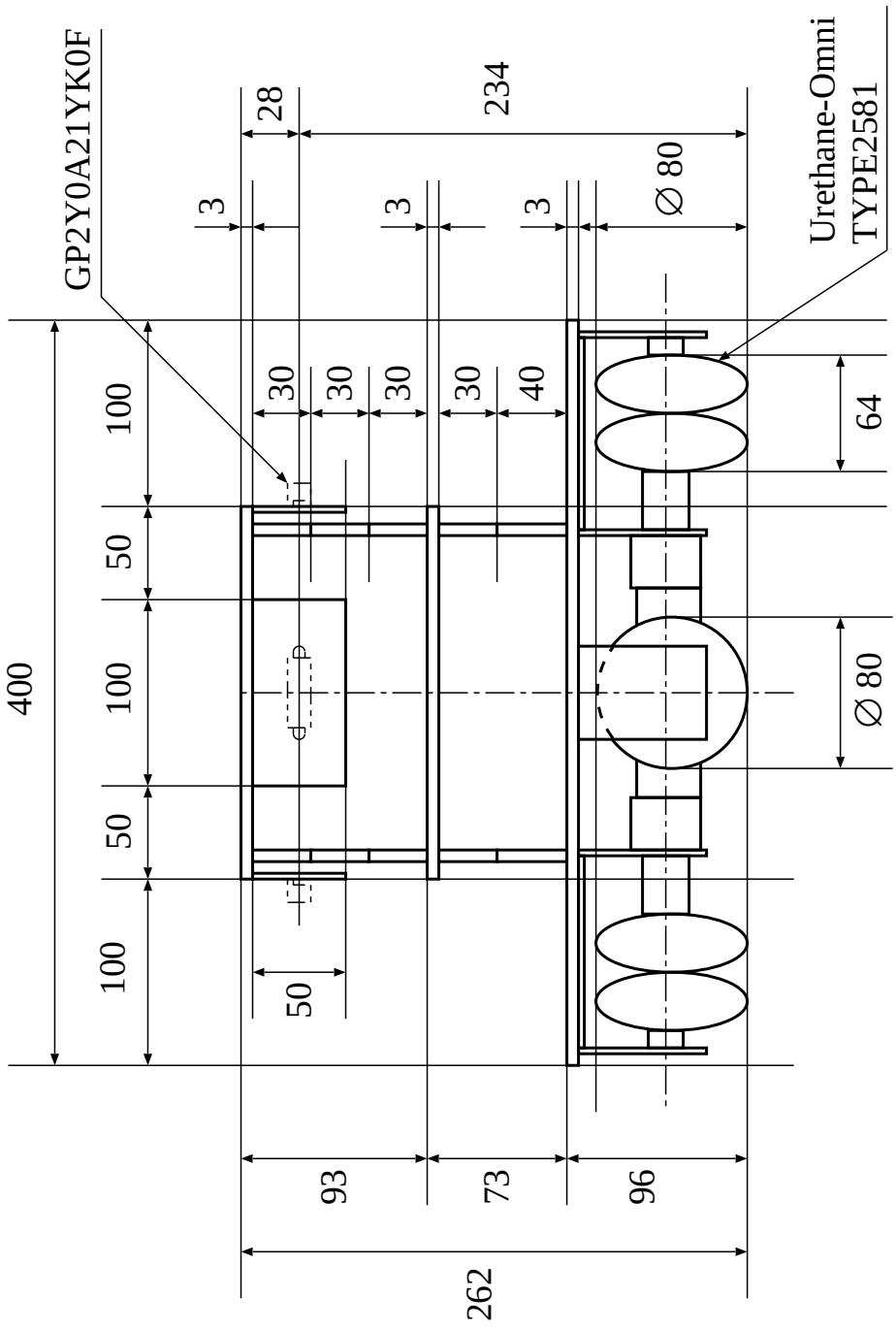


図 52: ロボットの正面図（反時計回りに 90 度回転）

A.3　回路図

本研究の実験で使用したロボットを構成する電子回路の回路図を記載する．ここでは，電子回路の構成を表現するブロック図と，H ブリッジ回路の実体配線図を記載する．

はじめに，ロボットの電子回路の構成を表現するブロック図を図 53 に示す．Armadillo-300 は，学習のプログラムを実行するプラットフォームとし

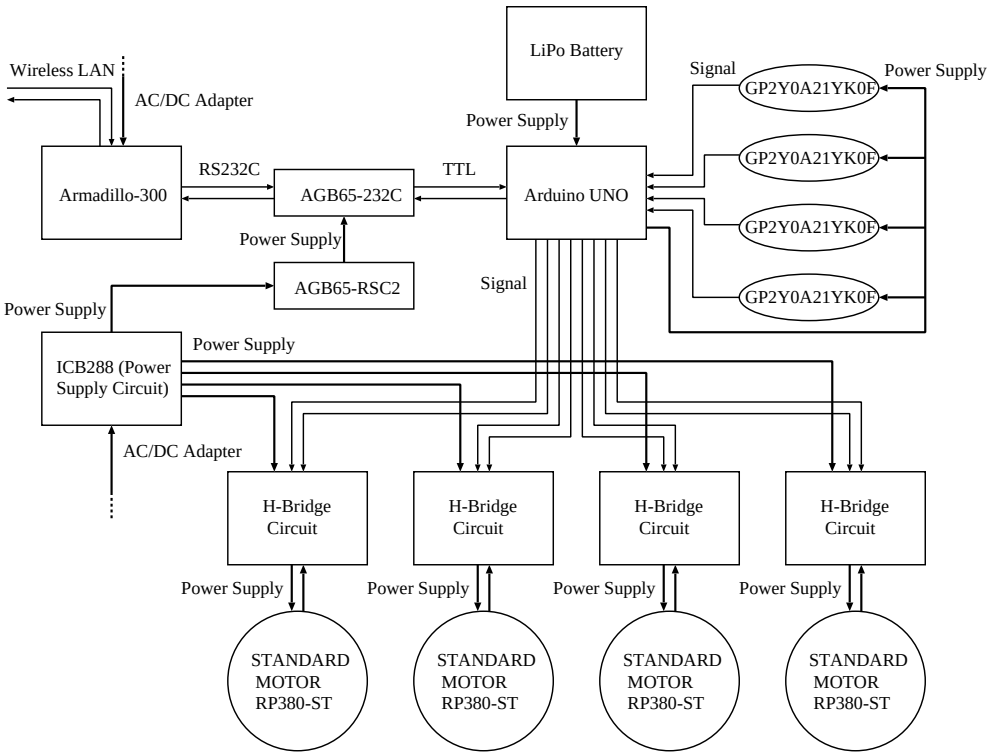
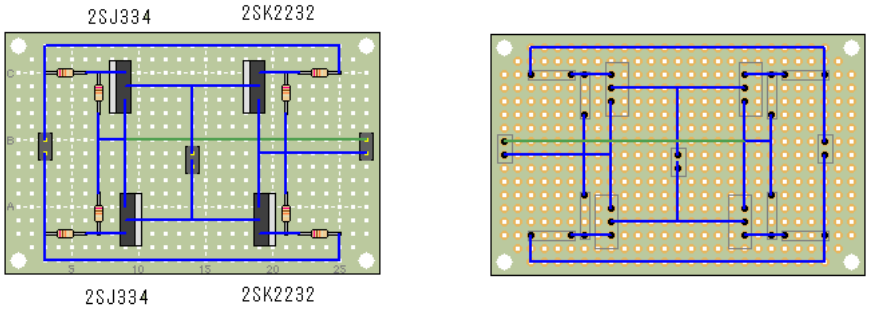


図 53: 電子回路の構成（ブロック図）

て動作し，Arduino UNO にモータ駆動の指令やセンサ情報の要求をする．AGB65-232C は，シリアル通信の信号レベルを変換するインターフェイス回路である．Armadillo-300 と Arduino UNO は，AGB65-232C を介してシリアル通信によって双方向にデータ転送する．このとき，Armadillo-300 は RS232C（PC）レベルの信号を出力し，Arduino UNO は TTL（5.0V マイコン）レベルの信号を出力する．Arduino UNO は，Armadillo-300 の命令に従って，モータを駆動させたりセンサ情報を取得する．Arduino UNO がモータを駆動させるときは，H ブリッジ回路に制御信号を送る．H ブリッジ回路は，2 値 × 2 本の制御信号の組み合わせにより，DC モータを任意の方向に回転させることができる．AGB-RSC2 は，RC サーボモータを使用するために搭載しているのではなく，単に AGB65-232C へ電力を仲介して供給

するように使用している．ICB288(Power Supply Circuit) は，オムニホイール・モータおよび AGB65-RSC2 へ電力を供給している．

次に，H ブリッジの実体配線図を図 54 に示す．ここで，図 54 は，プリント基板エディタ Paas (Parts Arrange Support System) を使用して作成したものである．緑色の線は裏面被膜配線，青色の線は裏面配線である．左側の



(a) 表面　　(b) 裏面

図 54: H ブリッジ回路の実体配線図

ピンソケットは制御信号用，真ん中のピンソケットはモータ駆動用，右側のピンソケットは電力供給用である．長方形の物体は，電界効果トランジスタ（FET）である．左側は p 型 FET であり，TOSHIBA 製の 2SJ334 である．右側は n 型 FET であり，同社製の 2SK2232 である．この H ブリッジ回路は，Arduino UNO から送信される制御信号に従って機能するものである．ここで，左側のピンソケットの上側が信号線 1，下側が信号線 2 とすると，これらの信号線の電圧の組み合わせとモータの回転は，表 6 の通りに対応する．これにより，Armadillo-300 は，Arduino UNO を介してモータを任意の方向に回転，もしくは静止やブレーキさせることができる．

表 6: H ブリッジ回路の機能

信号線 1	信号線 2	モータの回転
OFF	OFF	静止
OFF	ON	逆回転
ON	OFF	正回転
ON	ON	ブレーキ

参考文献

[1] 中村仁彦, Ole Jakob Sørdalen, Woojin Chung: ”非ホロノミック・マニピュレータの理論的設計と非線形制御”, 日本ロボット学会誌, July 1995, Vol.13, No.5, pp. 674-682

[2] 岡本球夫: ”医療福祉ロボットの開発と安全技術”, 日本ロボット学会誌, November 2011, Vol.29, No.9, pp. 770-772

[3] 石川和良, 青山元, 関淳也, 足立佳儀, 石村佐緒里, 薩見雄一, 向殿政男: ”清掃ロボットにおける安全技術とコンポーネント”, 日本ロボット学会誌, November 2011, Vol.29, No.9, pp. 773-776

[4] 若林潔, 菊地広行, 森利宏, 岡田隆光, 小山義郎: ”警備ロボットにおける安全部品”, 日本ロボット学会誌, November 2011, Vol.29, No.9, pp. 777-779

[5] 寺田一貴, 鈴木陽介, 長谷川浩章, 曽根聡史, 明愛国, 石川正俊, 下条誠: ”全方位検出・高速応答可能なネット状近接覚センサの開発”, 日本ロボット学会誌, November 2011, Vol.29, No.8, pp. 683-693

[6] 町野保, 南條義人, 柳原義正, 河田博昭, 岩城敏, 下倉健一郎, 武藤伸洋: ”カメラとプロジェクタを搭載した移動ロボットによる実空間視野共有型コラボレーションシステム”, 日本ロボット学会誌, July 2010, Vol.28, No.6, pp. 746-755

[7] RODNEY A. BROOKS, ANITA M. FLYNN.: ”FAST, CHEAP AND OUT OF CONTROL: A ROBOT INVASION OF THE SOLAR SYSTEM”
MIT Artificial Intelligence Lab, Cambridge, MA, USA.
URL: `people.csail.mit.edu/brooks/papers/fast-cheap.pdf`

[8] 久保田孝: ”探査機「はやぶさ」の AI 技術”, 人工知能学会誌, March 2011, Vol.26, No.2, pp. 156-163

[9] 久保田孝: ”惑星別探査ローバ”, 日本ロボット学会誌, July 2003, Vol.21, No.5, pp. 468-471

[10] 小田光茂, 久保田孝: ”日本の宇宙開発・宇宙探査の技術ロードマップ”, 日本ロボット学会誌, July 2009, Vol.27, No.5, pp. 482-489

[11] 大川一也, 茂垣彰人: ”離散的な獲得データに基づく自動駐車のための経路計画と連続した動作の実現”, 日本ロボット学会誌, May 2011, Vol.29, No.4, pp. 376-383

[12] 門根秀樹, 中村仁彦: ”自己組織的な非単調活性関数をもつ連想記憶モデルによる運動パターンの記号化とスパースコーディング”, 日本ロボット学会誌, November 2011, Vol.29, No.9, pp. 801-810

[13] 山口明彦, 高松淳, 小笠原司: ”強化学習によるロボットの動作獲得のための基底関数に基づく行動空間生成手法”, 日本ロボット学会誌, January 2011, Vol.29, No.1, pp. 55-66

[14] 岡本太一, 小林祐一, 大西正輝: ”ロボットの障害物回避行動生成における画像特長の獲得”，日本ロボット学会誌, December 2010, Vol.28, No.10, pp. 1213-1222

[15] 山口明彦, 杉本徳和, 川人光男: ”回避行動の再利用メカニズムを備えた強化学習手法と多関節ロボットの全身運動学習への応用”, 日本ロボット学会誌, March 2009, Vol.27, No.2, pp. 209-220

[16] 実川達明, 上田隆一, 新井民夫: ”サンプリング実時間 Q-MDP 法—不完全な観測に基づき実時間行動する自律サッカーロボットへの適用—”, 日本ロボット学会誌, January 2009, Vol.27, No.1, pp. 71-78

[17] 松井藤五郎, 後藤卓: ”強化学習を用いた金融市場取引戦略の獲得と分析”, 人工知能学会誌, May 2009, Vol.24, No.3, pp. 400-407

[18] 光永法明, 浅田稔: ”移動体の意思決定のための情報量基準に基づく観測対象選択戦略”, 日本ロボット学会誌, September 2001, Vol.19, No.6, pp. 793-800

[19] Richard S. Sutton and Andrew G.Bart (三上貞芳・皆川雅章 共訳): ”強化学習”, 森北出版株式会社，2001

[20] 新美智秀: ”センシング工学”, 株式会社コロナ社, 2004

[21] 森政弘, 小川鑛一, ”第 2 版 初めて学ぶ基礎制御工学”, 東京電機大学出版局, 2010

[22] 広瀬貞樹: ”あるごりずむ” , 近代科学社, 2009

[23] 和多田作一郎: ”AI の基礎を知る事典”, 実務教育出版, 1988

[24] 植草育三, 高谷芳明, 多根井文男, 深井完祐: ”初心者のための機械製図（第 2 版）”, 森北出版株式会社, 2009

[25] 日本工業標準調査会: ”知能ロボット - 用語” , Intelligent robots - Vocabulay , JIS B 0185:2002

[26] Pololu Robotics and Electronics: ”gp2y0a21yk0f.pdf” , http://www.pololu.com/file/download/gp2y0a21yk0f.pdf?file_id=0J85

[27] 株式会社 土佐電子: 軽量化モータフォルダ,
http://www.tosadenshi.co.jp/cargo/goodslist.cgi?in_kate=70-2

[28] 同上: 構成部材の図面, http://www.tosadenshi.co.jp/cargo/image/4mmmotfol.gif

[29] 株式会社 タミヤ: ギヤード・モータ図面（380 シリーズ） ,
http://www.tamiya.com/japan/robocon/robo_parts/g_motor/g_motor_zumen.htm

[30] 浅草ギ研: AGB65 シリーズ PC 接続ボード AGB65-232C の紹介 ,
http://www.robotsfx.com/robot/AGB65_232C.html

[31] 浅草ギ研: RC サーボコントローラ AGB65-RSC2 の紹介 ,
http://www.robotsfx.com/robot/AGB65_RSC2.html

謝辞

はじめに，本論文を作成するにあたり，日頃より懇切なる御指導を賜りました主指導教員の倉重健太郎先生に，深く感謝の意を表します．また，卒業研究中間発表会の場で，大変貴重な御指導と御助言，御意見を下さいました佐賀聡先生，畑中雅彦先生，本田泰先生に厚く御礼申し上げます．そして，本研究に関して多大な御協力を頂きました木島康隆さんに心より感謝致します．最後に，研究報告の場で貴重な御助言と御意見を頂きました認知ロボティクス研究室の宮崎愛央さん，中南義典さん，澁谷和さん，北山直樹さん，杉本大志さん，梅津祐介さん，三浦丈典さん，高泉昇太郎さんに感謝致します．