Vol.2012-EC-23 No.17
2012/3/27

ビデオゲームエージェントの自律的行動獲得と観測情報の信頼性に着目した獲得行動パターンの分析

佐藤　祐一[†1]　　片寄　晴弘[†1]

ビデオゲーム開発における課題の一つに，COM プレイヤ（ビデオゲームエージェント）の行動パターンの実装がある．その課題に際し，強化学習によってエージェントの行動パターンを自律的に獲得するアプローチの提案がなされているが，過度の最適化が進められる結果，強さが状況に応じて著しく揺らいだり，また，獲得された行動パターンが「機械的」と感じられるという問題があった．本研究では，Infinite Mario Bros を対象とし，観測情報に遅れや揺らぎが生じるという環境下での Q 学習による行動パターンの獲得を試み，上記の問題がどう解決されるかの検証する．

A study of Autonomous Acquisition of a Video-Game Agent Action under Imposing Physical Constraints.

Yuichi Sato[†1] and Haruhiro Katayose[†1]

One of the problems int he video game development has the implementation of the behavior pattern of the COM player (video game agent). On the occasion of the problem, suggestion of the approach to get the behavor pattern of the agent autonomously is made by reinforcement learning, but that has problem of a got behavior pattern ”is mechanical” cause of excessive optimization is pushed forward, and strength shakes depending on the situation remarkably, I carry out the inspection that the problem mentioned above are solved that the acquisition of the behavior pattern by the Q-Learning under the environment which is a delay and a fluctuation produce in observation information intented for ”Infinite Mario Bros”.

†1 関西学院大学理工学研究科情報科学専攻
Department of informatics, Graduate School of Science and Technology,Kwansei Gakuin University

1. 序　　論

1.1 研究背景

近年，ハードウェアの発展に支えられ，ビデオゲームのグラフィックやサウンドのクオリティが飛躍的に向上し続けている．技術の向上は，ゲーム世界におけるリアリティの向上を追求する価値観の醸成につながり，これに関連する形で，グラフィックやサウンドだけではなく，COM プレイヤ（ビデオゲームエージェント）の振る舞いや戦略自体のリアリティが重要視されるようになってきている[1)]．ビデオゲームエージェントは様々なタイプがあるが，代表的なものとしては，コンピュータ制御による強さの限界を求めるもの，対戦型ゲームなどにおいてプレイヤの対戦相手となりプレイヤに楽しさを提供するもの，プレイヤに対してデモプレイなどの形で参考となるプレイログを教示し，プレイヤに攻略法などを教えるものなどがあげられる．これらエージェントに求められる振る舞いのリアリティを確保することが求められている．

エージェントにおける「振る舞い」は，そのエージェントが置かれた状況，つまり，当該のゲームにおいて，様々なマップ上での相手プレイヤや敵キャラクタの状況を踏まえつつ，「高得点をあげたい」「安全策を講じたい」「敵をよけたい」などの状況に対応して，振る舞いのための行動パターンを規定されるべきものである．それをすべて手作業で設定しようとすると極めて煩雑な作業となる[2)3)]．それゆえ，この作業を学習アルゴリズムを用い自動化しようとする試みとして，経路探索問題に帰着する手法[4)5)]，人間のプレイデータからの学習による手法[6)7)]，強化学習に基づく手法[8)9)] が提案されてきた．

経路探索問題の代表的なアルゴリズムとしては A*アルゴリズムがあげられる．2009 年の Mario AI Competition[⋆1]では、，A*アルゴリズムに基づいて作られた Robin のプログラムが優勝した[5)]．この例が示すように，経路探索問題の解法を用いてきわめて強いエージェントを作ることも可能であるが，その振る舞いは人間離れしたものであり，人間プレイヤのための相手プレイヤエージェントを構成するという目的には適しているとは言えない．人間のプレイデータからの学習によるアプローチのうち代表的なものとしては，保木によって提案された将棋プログラム Bonanza があげられる．将棋のように強いプレイヤの膨大な棋譜データが用意できる場合には，自動学習による行動パターン獲得は有効である．一方，強

⋆1 設定されたアクションゲームをゲームをエージェントに攻略させ，優秀なエージェントを競う大会．2 章において詳しい説明を記述

化学習は，同じ学習に分類されるアプローチでも，膨大な事例を与える必要がなく，自身の「振る舞い」の試行を重ねて最適な行動パターンを獲得していく．このアプローチによる研究として[8)]〜[?)]があげられる．これらの手法では，強化学習がさまざまなジャンルのビデオゲームの行動パターン獲得に使用しうることを示している．

強化学習の枠組を用いれば，多くの対戦事例データを与えなくても行動パターンが自律的に獲得できるというメリットが存在する．しかしその一方で，極度の最適化が進む結果，早すぎる反応速度や正確すぎる行動制御，すなわち，人間にとって「機械的」と映る行動パターンが獲得されるという問題がある．また，局所的な条件において最強であっても，状況が少し変わっただけで，とたんに弱くなるタイプのエージェントが構成されてしまうという問題がある．これまでのビデオゲームを対象とし，強化学習を用いた研究では，獲得した行動パターンがプレイヤからみて「人間らしい」と感じる振る舞いを生成しているかという検討はほとんどなされてなかった．

本研究では，人間のプレイヤが,「人間のプレイヤがゲームをしている」ように感じられるエージェントの「振る舞い」の自律的に獲得しうる機構の構成を主題とする．具体的には，強化学習の枠組みにおいて，状況に応じて強さが著しく揺らいだり，また，獲得行動パターンが「機械的」と感じられるという問題を解決する方策の検討を目指す．本研究では，生物の身体的な制約の上で最適化され獲得された行動パターンが，人にとって「人間らしい」と映る，という仮説を立て，この仮説の実証実験を実施する．

人に限らず，生物であれば「見間違い」や「手が滑った」というようなセンサ系や運動系の誤りがあり，また，センサ系，制御系、センサ系の一連のプロセスにおいて「遅れ」が生じる．このような「ゆらぎ」「遅れ」を強化学習の学習系に組み込むことで，人間がプレイしているような感覚を提供するエージェントの「振る舞い」を構築できるかどうかを検証することが本研究の狙いである．加えて，経路探索問題に帰着する手法や，強化学習を使用した他の手法の，特定の条件では極めて強いが，少し状況設定が変ると極端に弱くなるという課題がどの程度軽減できるのかの検証を実施する．

1.2 関連研究

ビデオゲームを対象とし，強化学習を用いて行動パターン獲得を行った研究として[8)],[9)]があげられる．藤田らはカードゲーム Hearts を題材とし，複数のユーザーが参加するゲームにおいて Q 学習を用いて戦略獲得を行った．また，藤井らはポケットモンスターや遊戯王といったメジャーなトレーディングカードゲームを基にしたビデオトレーディングカードゲームを設定し，多層パーセプトロンを用いて戦略獲得を行った．このように，強化学習を用いた研究では，学習対象のビデオゲームに対する行動パターン獲得に成功している．しかし，学習の結果生まれた行動パターンが「人間らしい」振る舞いをしているかという議論はなされてなかった．

人間の持つ行動制御の特徴について研究しているものとして[10)],[11)]があげられる．Cabrera ら[10)] は人間の指先による直立棒の制御実験を行い，人間の行動制御の特徴を調査した．人間には情報処理能力の限界によって反応に遅れが存在し，物体の位置を観測するうえで誤差が生じてしまうといった身体的制約がある．また，直立棒の制御においては，直立棒の動きに大きなノイズが発生するため，完璧な予測ができないといった問題がある．しかし．この実験では，指先の動きの特徴的スケールが反応時間よりも短い場合が頻繁に観測され，また訓練により制御がよりうまく行うことができるようになるという実験結果が示された．この実験結果から，身体的制約を意識的もしくは無意識的に考慮し，指先の行動制御に対してノイズを取り入れているのではないかと提唱した．大平ら[11)] はノイズは人間の直立姿勢の制御においても有用であると示し，ロボットの直立姿勢制御に応用するための定式化を行った．これらの研究結果から，人間は身体的制約を考慮し，安全性とパフォーマンスを両立する特徴を持った行動パターンの獲得を行なっていると考えられる．本研究はこれらの研究結果に基づき，人間の持つ身体的制約をエージェントに組み込むことにより，行動パターンにどのような変化が生じるか，人間による感じ方に変化が生じるかについて検討を行うこと目的としている．

2. 問題設定

2.1 本研究における課題

本研究は「ゆらぎ」や「遅れ」といった人間の持つ身体的制約を強化学習エージェントに組み込むことで,「人間らしい」と感じる行動パターンの獲得を目指す．そのためには次の課題を解決する必要がある．

- **状況に応じた性能のゆらぎの解決**

  本研究ではビデオゲームという理想環境で動作するエージェントに対し，ゆらぎや遅れにより観測情報が必ずしも信頼できる情報ではない非理想環境での学習を目指す．そのためにはゆらぎや遅れを強化学習に組み込むためのパラメータ化が必要となる．また，理想環境ではエージェントは正確な情報を得ることができ正確な行動制御を行うことができる．しかし，観測情報に変化が発生した場合，正確な情報を得ることができないため正確な行動制御が難しくなり，エージェントの性能に大きなゆらぎがでてしま

う．本研究では観測情報がゆらぎや遅れが発生する環境において行動パターンを獲得する手法の提案も行う．

- **ビデオゲームにおける強化学習アルゴリズムの決定**
  強化学習の枠組みにおいてどのアルゴリズムを用いるか決定しなければならない．ビデオゲームにおいては，環境が完全な既知ではなく，また報酬分布のモデルが与えられていないため，それらの知識を持たず学習する必要がある．また，ビデオゲームではある状態においてとった行動に対して直接的な評価をしなければならない．これはビデオゲームにおいては状態が重要なのではなく，あくまである状態において取った行動がどれほど有効であるかを評価する必要があるからである．本研究ではそのような知識を必要とせず，行動に対して直接評価を与え，学習を行うことのできる強化学習の枠組みとして Q 学習を用いて学習を行う．
- **学習対象の決定**
  本研究はビデオゲームにおいて行動パターンを獲得し，獲得行動パターンの比較を目的としている．そのためには限りなく同じ状況を再現できるビデオゲームが望ましい．また,「ゆらぎ」や「遅れ」が発生し，獲得行動パターンに影響がでるのはリアルタイム制御が必要なゲームある．最後に，明確な学習目標を設定できることが重要である．強化学習において行動パターンの獲得を行うためには，学習目標が必須である．これらを容易に実現できる学習対象として，本研究では Inifinite Mario Bros[12] を用いる．

2.2 Infinite Mario Bros

本研究では学習対象として，本研究では Inifinite Mario Bros[12] を用いる．Infinite Mario Bros は，ランダムに生成されるマリオライクなステージを制限時間中攻略するアクションゲームである．ステージの自動生成は事前に与えたシード値にしたがって行われる．

2.2.1 Infinite Mario Bros の特徴

Infinite Mario Bros における特徴・仕様は以下のようになっている．

- **エージェントの操作キャラクタ（マリオ）**
  Infinite Mario Bros ではエージェントはマリオを操作する．エージェントによるマリオの操作はキー入力 (LEFT,RIGHT,DOWN,SPEED,JUMP) を用いて行う．フレーム毎のそれぞれのキーの押下状態により，マリオは対応した行動を行う．また，マリオの状態として”大”状態・”小”状態が存在する．この状態は後述する被ダメージ条件を満たすことで変化する．
- **敵キャラクタ**
  Inifinite Mario Bros では複数種類の敵キャラクタが登場する．敵キャラクタはそれぞれ独自のアルゴリズムで動作している．この敵キャラクタはいわゆる「お邪魔キャラ」として設定されており，エージェントはこの敵キャラクタを避けて進むか，倒して進むかなどどのように処理するかが求められる．
  マリオは敵キャラクタの接触判定によってダメージを受ける．接触判定について，踏むことができる敵キャラクタについては踏む以外の行動で接触した場合ダメージを受ける．踏むことができない敵キャラクタについてはどのような行動で接触した場合についてもダメージを受ける．
- **スコアの獲得**
  マリオが死亡するまたは設定された制限時間に達すると攻略は終了し，スコアを獲得する．スコアは既定の評価関数で計算され，敵キャラクタを倒した数，そしてステージを攻略した距離などに応じてスコアが上昇する．獲得スコアが高いほど優秀なエージェントとして評価することができる．
- **エージェントの観測情報**
  エージェントは様々な情報を観測情報として得ることができる．観測情報として，マリオの座標，マリオの状態，画面内の敵キャラクタの種類および座標，地形情報といったものを観測情報として得ることができる．エージェントの観測する地形情報は，ステージに配置されているブロックのうち，画面内にある 22*22 のブロックの配置情報となる．これらがエージェントの観測情報として毎フレーム与えられる．Infinite Mario Bros は 1 秒 24 フレームで動作しており，エージェントは毎フレーム観測情報を受け取りマリオの行動制御を行うためのキー入力を返す必要がある．

2.2.2 Mario AI Competition

Infinite Mario Bros を対象としたエージェント評価コンテストとして Mario AI Competition が開催されている[13]．評価方法は攻略によって獲得するスコアの高さを評価軸としている．代表的なエージェントとして，Robin が開発したエージェント[5] がある．敵キャラクタの動きやマリオの動きを事前に学習・解析したうえで，A*アルゴリズムを用いたルート探索によって攻略をしている．また，スコア獲得方針は「敵キャラクタは可能な限り避け，とにかく早くステージをより遠くまで攻略し高いスコアを獲得する」としている．

Infinite Mario Bros において学習目標の設定は重要であり，獲得行動パターンにも直接影響する．Robin のエージェントは多くのエージェントが寄せられている Mario AI Competition において優勝したエージェントであり，攻略における 1 つの最適解として扱

うことができると考えられる．そこで，本研究における強化学習を Robin のエージェントのものと同一のものに設定する．

3. 観測情報の変化に対応する行動パターン獲得

本章では,「ゆらぎ」や「遅れ」といった制約を持ったエージェントが行動パターンを獲得するための手法について述べる．まず「ゆらぎ」や「遅れ」をエージェントに落とし込むためのパラメータ化として観測情報の信頼性の定義を行い，パラメータ化により変化する観測情報に対応する手法を述べ，学習する手法について述べていく．

3.1 観測情報の信頼性

人間の持つ身体的制約による「ゆらぎ」や「遅れ」を強化学習に付与するためのパラメータ化として,「観測情報の信頼性」を定義する．観測情報の信頼性とは，観測情報の誤差および観測情報の遅延により観測情報がどのように変化するかという変化の大きさを表すものである．観測情報の信頼性として,「情報認識の遅れ」および「観測位置情報の誤差」の 2 点を定義する．

プレイヤは情報を観測から行動を行う過程で，得られた情報を認識し，実際に行動をするまでにタイムラグが発生する．つまり実際に行動する時点は観測情報は過去の情報となる．これを情報認識の遅れと定義し，ビデオゲーム側からエージェントに渡す観測情報を本来を遅らせることで再現する．エージェントに渡す観測情報を遅らせるほど，エージェントは過去の情報を認識することになる．

非理想環境において，観測したオブジェクトの正確な位置（座標）の認識は難しく観測位置情報の誤差が発生する．この誤差は観測したオブジェクトに対する位置ノイズとして発生する．この位置ノイズをエージェントの観測情報における観測位置情報の誤差として定義し，エージェントが観測したオブジェクトの座標に対してノイズを付与することで再現する．Infinite Mario Bros に対しては，観測したマリオの座標および敵キャラクタの座標に対してノイズを付与することで本来の座標とは誤差のある座標をエージェントは観測する．本研究ではこの位置ノイズをガウス分布に従って発生すると仮定し，ガウス分布の分散値が大きくなるほど誤差は大きくなる．

定義したこれら 2 点について，観測情報の遅れの大きさおよびガウス分布の分散値の大きさを観測情報の信頼性のパラメータとして扱う．定義した 2 点のパラメータを操作することで，観測情報に対する信頼性を変化させる．

3.1.1 観測情報の変化への対応

本節では定義した観測情報の信頼性パラメータによって変化した観測情報に対応する手法について述べる．非理想環境における観測情報は，真の観測情報を比較すると現在の情報ではなく過去の情報を観測し，またマリオおよび敵キャラクタの認識した座標にガウス分布に従ったノイズが付与されており観測情報が変化している．そのため観測情報の誤差を小さくし，観測情報の信頼性を向上しなければならない．そこで，得られた観測情報から真の観測情報を予測することで観測情報の信頼性を向上を行い，学習における影響の低減をはかる．

真の観測情報を予測するためにはマリオおよび敵キャラクタの真の座標の予測を行う必要がある．真の座標を予測するため，過去に行った移動パターンに基づいた移動予測を行う．移動パターンの抽出のためには，あるフレームにおいてマリオおよび敵キャラクタがどのような状況下でどのような移動パターンを行ったかをセットで抽出しなければならない．しかし，観測情報の全てを利用した場合十分予測可能な情報が集まるまでに非常に多くのログが必要となってしまうため，観測情報を圧縮することで状態数の圧縮を行う．敵キャラクタについて，次のように情報の圧縮を行った．

- 対象の敵キャラクタを中心とした上下左右方向それぞれの障害物ブロックまでの距離（ブロック単位）
- マリオとの x 軸および y 軸方向の距離（ブロック単位）
- 対象の敵キャラクタの進行方向

次に，マリオについて次のように情報圧縮を行った．

- マリオを中心とした周囲 8 ブロックの地形情報
- マリオを中心とした周囲 8 ブロックの敵キャラクタの配置
- 前フレームのキー入力

以上のように圧縮した情報はそのフレームにおける瞬間的な情報しか含まれておらず，連続的な時間を含んだ情報であることがより正確な移動パターンの抽出のために望ましい．そこで，対象のフレームから過去の連続したフレームを系列化して抽出することで，連続的な変化として扱うことができ，より正確な移動パターンの抽出をおこなうことができる．

3.2 Q 学習による行動パターン獲得

Infinite Mario Bros において行動パターンを獲得するための Q 学習について述べる．

3.2.1 Q 学習概要

Q 学習とはある状態を s, エージェントがとった行動を a とした場合，状態 s と行動 a

を組とし，その組に対する Q 値とよばれる評価値を１つの”ルール”として設定する．Q 値は全てのルールで独立して存在し，Q 値が高いほどそのルールが有効であるといえる．

Q 学習における Q 値の更新式は以下のものを用いる．

$$Q(s_t, a_t) = (1-\alpha)Q(s_t, a_t) + \alpha((r + \gamma max_p Q(s_{t+1}, p)) \quad (1)$$

数式 1 において，t はフレーム，s_t はフレーム t における状態，a_t はフレーム t においてとった行動，$Q(s_t, a_t)$ は (s_t, a_t) に対応するルールの Q 値である．α は学習率呼ばれ，Q 値の更新において新たな報酬をどれだけ重視するかを示す値であり，γ は割引率と呼ばれる数値である．r は状態 s_t における行動 a_t をとったことによって得られる報酬である．本研究では行動選択手法として ε -greedy 法を用いる．

3.2.2 状態と行動の設定

Q 学習は学習時間が状態数の指数関数オーダーであるため，行動パターンを獲得することができ，かつ現実的な学習時間で収束のできる状態を設定する必要がある．そこで，観測情報を圧縮し状態を次のように設定する．

- **マリオを中心とした 7*7 ブロックの地形情報**

  エージェントが観測する地形情報は画面を 22*22 ブロックに分割したものである．しかし，1 フレームあたりにエージェントの操作対象となるマリオの移動距離は小さく，画面内全ての地形情報がマリオの動作に影響することはほとんどない．そこで，マリオの行動に影響がでる範囲を考慮し，地形情報に関する状態をマリオを中心とした 7*7 ブロックの地形情報に圧縮する．

- **マリオを中心とした 7*7 ブロックの敵キャラクタの配置**

  地形情報と同じように，１フレームあたりのマリオの移動距離は小さいため，画面内全ての敵キャラクタの情報は必要なく，マリオに近い敵キャラクタの位置情報が重要である．また，観測情報における敵キャラクタの位置は座標で与えられるが，座標単位で状態設定した場合状態数の肥大化に繋がる．そこで，地形情報と同じようにブロック単位での位置情報に圧縮し，マリオを中心とした 7*7 のブロック座標での位置情報に圧縮した．

- **マリオの状態**

  マリオの行動に対してマリオの状態は大きく影響しない．しかし，大状態でダメージを受けた場合は小状態に変化するだけで攻略を続行できるが，小状態でダメージを受けた場合は死亡扱いになるため，攻略をより長く進めるうえでマリオの状態を Q 学習における状態として設定することは必要であると考えられる．

- **マリオの進行方向**

  Infinite Mario Bros においてマリオの進行方向は重要である．例えば右に敵キャラクタがいる状況を想定した場合，マリオが右に移動している場合はその敵キャラクタを考慮した動きをとる必要がある．しかし，マリオが左に移動している場合は敵キャラクタから離れていく動きとなるため，その敵キャラクタを考慮する必要性は少ない．ここから，マリオの進行方向を８方向＋停止の９状態としたものをマリオの進行方向として設定する．

次に，行動の設定について述べる．マリオの制御はキー入力のよって行う．このキー入力の組み合わせに対して，行動制御において影響のある９つの組み合わせを可能な行動として設定する．設定した全ての可能な行動を表 1 に示す．

3.2.3 報酬の設定

Q 学習の報酬の設定について述べる．本研究では「敵キャラクタを可能な限り避け，とにかく遠くまでステージを攻略する」と学習目標を設定した．この目標では早く進むことに対して正の報酬を与え，逆にダメージを受ける，死亡するといった攻略を阻害する要素に対して罰則を与えることが望ましい．そこで，目標に合わせ報酬 reward を次のように設定した．

$$reward = distance + dameged + death \quad (2)$$

distance は使用したルールによって進んだ距離をそのまま報酬とする．dameged は使用したルールを用いてダメージを受けた場合に与える負の報酬，death は使用したルールを用いて死亡した場合に与える負の報酬である．報酬の調整の結果，damaged は-50，death は-100 を採用した．

4. 実験と考察

本章では行動パターン獲得手法を用いて，実際に行動パターンを獲得することができたかについて検証を行う．次に，観測情報の変化により獲得行動パターンにどのような違いが生まれたかについて検証する．最後に，獲得した行動パターンが人間にどの程度人間らしい動作として見えるかについて，初期的検討を実施する．

4.1 エージェントの学習性能の検証

提案した行動パターン獲得手法が有効であることを示すため，観測情報の信頼性パラメータを複数与えた場合における獲得スコアの推移を調べる．

表 1 行動の種類とキー入力の組み合わせ

行動の種類	(LEFT,RIGHT,DOWN,JUMP,SPEED)
右に歩く	(OFF,ON,OFF,OFF,OFF)
右に走る	(OFF,ON,OFF,OFF,ON)
右に歩きながらジャンプ	(OFF,ON,OFF,ON,OFF)
右に走りながらジャンプ	(OFF,ON,OFF,ON,ON)
左に歩く	(ON,OFF,OFF,OFF,OFF)
左に走る	(ON,OFF,OFF,OFF,ON)
左に歩きながらジャンプ	(ON,OFF,OFF,ON,OFF)
左に走りながらジャンプ	(ON,OFF,OFF,ON,ON)
しゃがむ	(OFF,OFF,ON,OFF,OFF)

獲得スコアの推移を調べるため，シード値をランダムに与え，毎回新たにランダム生成されるステージを対象として学習試行を行う．試行回数は 10 万回とし，200 ゲームごとの獲得スコアの平均をとる．学習のための Q 学習に関連するパラメータ設定として，学習率 α を 0.2，割引率 γ を 0.9，ε greedy 法におけるランダム選択確率を 0.05 と設定した．エージェントに与える観測情報の信頼性パラメータとして，遅れ 6 フレーム (約 0,25 秒)・分散 8(半ブロック分相当)，遅れ 6 フレーム (約 0.5 秒)・分散 8(1 ブロック分相当)，および理想環境に相当する遅れ 0 フレーム・分散 0 という 3 つのパラメータを用いた．また，獲得スコアの比較対象として 2 章で紹介した Robin のエージェントを用いた．その結果を図 1 に示す．図 1 から，それぞれの観測情報の信頼性パラメータにおいても学習を進んでいることが確認できた．理想環境に相当する信頼性パラメータを与えたエージェントの獲得スコアについて，理想スコアに相当する A*エージェントの獲得スコアに近いスコアを獲得できていることから，本手法が学習目標に対して効率のよい行動パターンを構築できていることがわかる．

次に，観測情報の変化に対応する手法が，様々な環境の変化に対して適切に対応できているか検証する実験を行う．あらかじめ観測情報の信頼性パラメータを与え行動パターンを獲得したエージェントに対して，新たに様々な信頼性パラメータを与えて獲得スコアがどのように変化するか検証する．実験に使用する Q 学習に関連するパラメータおよび観測情報の信頼性パラメータは先程と同じものを用いた．比較対象として観測情報の変化への対応手法を持たせていない Q 学習のみを用いたエージェントにも信頼性パラメータを持たせて動作させる．学習試行回数は 10 万回とし，10 万回試行後学習をストップさせ，様々な信頼性パラメータを与え 200 試行の獲得スコアの平均をとった．その結果を図 2 に示す．図 2 において，観測情報の変化への対応手法を持たせていない学習エージェントは，信頼性パラメータによって観測情報の変化が大きくなるにつれ，大きく性能が低下していることがわかる．一方，観測情報の変化への対応手法を持たせたエージェントは観測情報の変化が大きくなるにつれ，スコアの減少は見られるが，対応手法を持たせていないエージェントと比べると減少の幅は小さい．この結果から，学習に用いた環境から更に環境が変化した場合でも性能のゆらぎを小さく抑えることができたことがわかり，状況の変化に弱いという強化学習の問題を低減できたことがわかる．

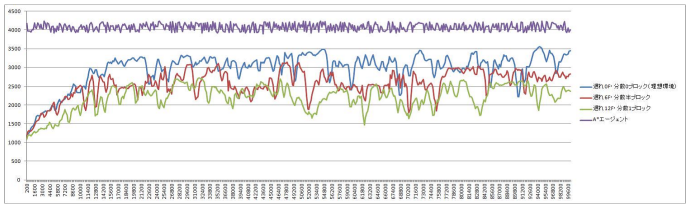

図 1 観測情報の信頼性パラメータを与え学習し，獲得したスコアの推移表

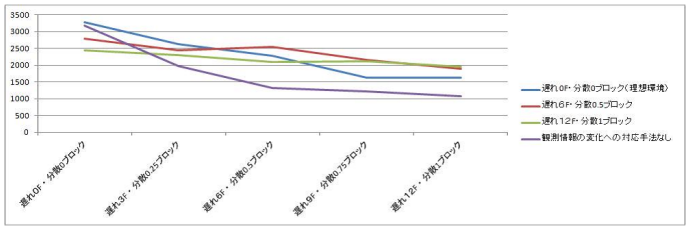

図 2 観測情報の信頼性パラメータを変化させた場合における獲得スコアの推移表

4.2 行動パターンの検証

信頼性パラメータを与え獲得した行動パターンが，理想環境で獲得した行動パターンと比べ行動パターンの特徴がどのように変化したか検証を行う．理想環境に相当するパラメータと非理想環境 (遅れ 6 フレーム・分散 8) のパラメータを与え学習したエージェントの行動パターンを比較する．それぞれの最適な行動パターンを比較対象とするため，各施行で同じシード値を与え，毎回同じステージが生成されるようにした状態で 10 万回学習試行を行い，試行中最も高いスコアを獲得したプレイログを比較する．学習のための Q 学習に関連するパラメータ設定は 4.1 節と同じものを用いた．

比較の結果，現れた行動パターンの傾向の違いについて図 3 に示す．図 3 上は倒すことができず，触れるだけでダメージを受けてしまう敵の攻略における行動パターンの違いである．理想環境では最小限のジャンプでノンストップで攻略しているのに対し，非理想環境では大きくジャンプを取り，また途中一瞬止まるような動作をしつつ攻略した．次に，図 3 中は大量の敵キャラクタが存在する区間の攻略における行動パターンの違いである．理想環境では信頼できる観測情報により，正確な行動制御を持って敵キャラクタが大量に存在する区間に突入しているのに対し，非理想環境では区間の手前で待機し，安全にいける状態に変化するのを待ってから突入するといった行動パターンが見られた．最後に，図 3 下は落ちると死亡する穴に対する攻略における行動パターンの違いである．理想環境では穴に落ちる寸前のところで最小限のジャンプで攻略しているのに対し，非理想環境では穴の少し手前で大きくジャンプをし，余裕を持って攻略を行なっていた．理想環境ではパフォーマンスのみを重視しており，非理想環境では安全性も考慮した行動パターンを獲得している．このような特徴の違いから，身体的制約を強化学習に組み込むことで行動パターンの特徴が変化していることがわかった．

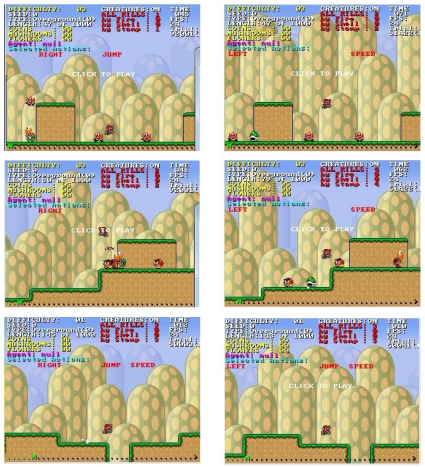

図 3 理想環境での獲得行動パターン（左）と非理想環境での獲得行動パターンの（右）比較画像

4.3 人間らしい行動パターン獲得の初期的検討

提案手法によって獲得した行動パターンがどの程度「人間らしさ」を持ったかについて，初期的検討を行う．初期的検討として 4.2 節で用いた 2 つの信頼性パラメータのエージェント，そして人間のプレイログを被験者 6 名による視聴実験によって人間らしさを判定した．実験用プレイログを取得するため，エージェントに対しては 4.1 節と同じ Q 学習に関連するパラメータを用い，4.2 節と同じく学習対象のステージに対して 10 万回学習試行を行い，試行中最も高いスコアを獲得したプレイログを抽出した．また，人間のプレイログを取得するため，プレイログ取得のためのログ抽出対象者を用意し，対象のステージを 15 分間プレイしてもらい，最もスコアの高かったプレイログを抽出した．

獲得したそれぞれのプレイログを「敵キャラクタが存在する」「穴が存在する」といったステージ中の特定の区間で切り取った動画をそれぞれ 6 動画ずつ作成した．比較のための手法として，シェッフェの一対比較法の浦の変法を用いた．それぞれの動画について，同じ区間を対象とした 2 動画を被験者に視聴してもらい，どちらが人間らしかったかを 4 段階で点数化し判定してもらった．

実験の結果，非理想環境で動作するエージェントと理想環境で動作するエージェントとの間に有意差または有意傾向があった区間について示す．1 区間について非理想環境で動作するエージェントが理想環境で動作するエージェントよりも人間らしいという有意傾向が見られた（有意水準 0.1）．また，人間の動作が理想環境で動作するエージェントよりも人間らしいという有意傾向も見られている．この区間では，触れることのできない敵キャラクタの存在する区間である．理想環境で動作するエージェントは最小限のジャンプを行い，ノンストップで攻略するという特徴を持った行動パターンであったが，人間の動作および非理想環境で動作するエージェントは，敵キャラクタに対して大きなジャンプを行い，また一瞬止まるような特徴を持った行動パターンであった．

別の 1 区間について，先程とは逆に，人間の動作および理想環境で動作するエージェントの両方が，非理想環境で動作するエージェントよりも人間らしいという有意差が見られた（有意水準 0.05）．この区間では，高台に存在する 5 体の敵キャラクタを攻略する区間である．この区間では，非理想環境で動作するエージェントは敵キャラクタを前に左右に細かく動くノイズのような動作が見られた．

4.4 考　　察

実験結果において，1 区間において非理想環境で動作するエージェントが理想環境で動作するエージェントよりも人間らしいという有意傾向が見られた．この区間における行動

パターンの特徴として，理想環境で動作するエージェントは最小限のジャンプを行い，ノンストップで攻略するというような特徴，つまりパフォーマンスを重視する特徴であったが，人間の動作および非理想環境で動作するエージェントは，敵キャラクタに対して大きなジャンプを行い，また一瞬止まるような特徴があった．この特徴の違いから，安全性を考慮した行動パターンが人間らしいと感じる要因である可能性を示すことができると考えられる．また，実験の際行った「何を基準に人間らしいと判断したか」という自由記述質問において，「敵キャラクタの前で一瞬ブレーキをかける」,「ためらいがある」といった回答があり，人間らしいという判断基準の１つとして考えられているところが見られた．

理想環境で動作するエージェントが非理想環境で動作するエージェントよりも人間らしいという有意差が見られた，先程とは逆の結果が現れた原因として，行動パターンに含まれるノイズが考えられる．この区間では，非理想環境で動作するエージェントは小刻みに左右に動きながらジャンプするという動作であり，明らかにノイズが大きく含まれている動作になっていた．このようなノイズが「人間らしくない」と被験者は感じたことが原因であると考えられる．自由記述質問においても「なめらかに動いている」,「無駄にジャンプをしない」を人間らしいという判断基準にしたという回答があった．この質問からも行動制御に含まれるノイズが人間らしさを損なっているということが考えられる．このような結果から，行動制御に含まれるノイズを抑えるための新たな身体的制約，例えば「疲れ」といった制約の導入が必要であると考えられる．

ビデオゲームにおいて「人間らしさ」はジャンルごと，ゲームタイトルごとに違った要素が考えられ，通常「人間らしさ」を組み込む場合様々な人間らしさの解析が必要となる．しかし，本研究では解析を行わず「身体的制約」のみを組み込むことで人間らしさを持つ可能性を示した．そのため，解析を行わず人間らしさを表すことができ，様々なジャンルに対して有効であるのではないかと考えられる．

5. まとめと今後の課題

本稿では，ビデオゲームを対象とし，強化学習により獲得した行動パターンが「機械的」と感じられるという問題を解決するため，人間のもつ観測情報に関する制約をエージェントに持たせることにより，人間の行動パターンの特徴である安全性とパフォーマンスを両立した行動パターンの獲得を目的とし，人間の持つ身体的制約を導入することを提案し実装を行った．提案手法を用いた視聴実験では，獲得行動パターンがどの程度「人間らしさ」を獲得したかについて，被験者による一対比較法による初期的検討を行った．この実験から，人間らしさ表す有意傾向が見られ，人間らしい振る舞いが観測できた．これにより，本研究の枠組みを用いることで，自律的に行動パターンを獲得しつつも，人間らしい振る舞いを持たせることができる可能性を提示した．今後の課題として，人間らしくないと感じる原因である行動制御に含まれるノイズを抑える枠組みが必要となる．現在身体的制約として「ゆらぎ」および「遅れ」の２つを強化学習の枠組みに加えているが，例えば「疲れ」といったような余分な動作を抑える身体的制約を加えることで抑えることができるのではないかと考えられる．今後はこのような行動制御のノイズを抑える身体的制約の提案および実装を行う．

参 考 文 献

1) 三宅陽一郎：ディジタルゲームにおける人工知能技術の応用，人工知能学会誌 23 巻 1 号 (2008)
2) 三宅陽一郎，横山貴規，北崎雄之: エージェント・アーキテクチャに基づくキャラクタ AI の実装，第 4 回デジタルコンテンツシンポジウム講演予稿集 (2008)
3) Orkin.J: 3 States and a Plan:The AI of F.E.A.R, Game Developer's Conference Proceedings(2006)
4) Slawomir Bojarski, Clare Bates Congdon：REALM: A Rule-Based Evolutionary Computation Agent tha Learns to Play Mario, 2010 IEEE Conference on Computational Intelligence and Fames(CIG'10)
5) Robin Baumgarten, Infinite Mario Bros AI[Online] http://www.doc.ic.ac.uk/rb1006/projects:marioai
6) 保木：局面評価の学習を目指した探索結果の最適制御. 第 11 回ゲームプログラミングワークショップ, pp. 78.83, Nov. 2006.
7) 星野准一, 田中彰人, 濱名克季: 模倣学習により成長する格闘ゲームキャラクタ, 情報処理学会論文誌 Vol.49 No.7(2008)
8) 藤田，石井信：マルチエージェントカードゲームのための強化学習法の改良，電子情報通信学会技術研究報告,Vol.102, No.731, pp.167-172(2003)
9) 藤井寂人，片寄晴弘：戦略型トレーディングカードゲームのための戦略獲得手法，情報処理学会論文誌, Vol.50, No.12 2796-2806(Dec.2009)
10) J.L.Cabrera,J.G.Milton:On-Off Intermittency in a Human Balancing Task, Physical Review Letters,89158702(2002)
11) 大平徹，保坂忠明：不安定な状況でのノイズと遅れの役割と制御への考察，交通流のシミュレーションシンポジウム，pp19-22(2004)
12) Inifinite Mario Bros.[Online] http://www.mojang.com/notch/mario/index.html
13) J.Togelius,S.Karakovskiy,J.Koutnik,J.Schmidhuber: Super Mario Evolution, CIG'09 Proceedings of the 5th international conference, IEEE press,2009, pp.156-161