Vol.2022-CG-188 No.26
Vol.2022-DCC-32 No.26
Vol.2022-CVIM-231 No.26
2022/11/19

身体性と内発的動機を伴う強化学習エージェントにおける描画行動の解析

阿部 由吾[1,a)]　米倉 将吾[1]　大村 吉幸[1]　國吉 康夫[1]

概要：近年，機械学習技術を用いた画像生成の分野は急速に発展している．多くの場合は，人間が用意したデータセットに基づく教師あり学習によって，システムがどのように画像を構成すべきかを学習する．同様に，提示された画像の美的評価を行うモデルが複数提案されているが，これらも人間があらかじめ美的評価を行った画像のデータセットに基づく教師あり学習で訓練されたものである．一方で，人間の美的感覚は，正解を外から教え込まれるのではなく，経験を通して内発的に獲得されるという側面がある．筆者らは，自発的な描画の学習を通して，エージェントにとっての美的評価基準が徐々に形成されていく過程をモデル化し，美的感覚の発達過程を構成論的に理解することを目指している．本稿では，身体性を伴った強化学習エージェントが，内発的動機によってどのような描画行動を示すのかを調査・解析した結果について報告する．

Drawing behaviors of reinforcement learning agent with embodiment and intrinsic motivation

1. 研究背景と目的

人工知能（Artificial Intelligence, AI）の分野において，近年，機械学習手法が大きく発展し，データ駆動の教師あり学習手法によって，機械が人間の美的評価行動を予測することができるようになった [1]．また，画像生成や音楽生成の領域も大きく発展し，機械がアート作品のようなものを創作する事例も報告されている [2]．しかしながら，AIが独自の美的評価基準を持ち，自律的に創作活動を行う事例は未だ存在していない [3]．

一方，人間の美的感覚は，正解を外から教え込まれるのではなく，経験を通して内発的に獲得されるという側面がある．美的体験における神経メカニズムを研究する神経美学によれば，人が知覚する美のカテゴリーには，衣食住などの生理的欲求に直結した生得的コンセプトの美と，アート作品・道徳・数学などの内的・社会的報酬に直結した後天的コンセプトの美の二つがあるという [4]．現在の神経美学において，なぜ人は，生理的欲求を満たすとも限らないアート作品や道徳などの後天的コンセプトの美を知覚するのかということが重要な問題として挙げられている．このような経験を通した美の知覚の形成は発達の問題として捉えることができ，本研究では，知能の発達原理の解明のために採られる手法である構成論的アプローチ [5] を採用して，この問題を扱っていく．発達研究における構成論的アプローチは，発達科学の知見に基づいた基本原理を想定し，身体を持ったエージェントに埋め込んで，環境との相互作用の中から現れる振る舞いを観測し，それを人間の振る舞いと比較することで想定した基本原理の妥当性を検証し，改善を加えるという研究手法である．

エージェントの美の知覚が経験に伴って発達するという仮説を調査するため，本研究では，多岐にわたる美的体験のシナリオの中から，特にアート創作・鑑賞のプロセスに焦点を当てる．アート創作・鑑賞のプロセスは，美的体験が関係するシナリオの中で最も一般的なものであり，心理学や神経科学，人文科学における数多くの知見が存在する．アート鑑賞における美的体験は，複数の段階の認知的情報処理と感情状態の相互作用の結果，対象の美的性質を判断する美的判断と，対象を受けて知覚される感情体験の二つの側面からなり，これらはそれまでの経験や専門的知識に依存するものだと言われれている [6]．これを踏まえ，本研

[1] 東京大学
The university of Tokyo, Bunkyo, Tokyo, 113–8656, Japan
[a)] y-abe@isi.imi.i.u-tokyo.ac.jp

ⓒ 2022 Information Processing Society of Japan

Vol.2022-CG-188 No.26
Vol.2022-DCC-32 No.26
Vol.2022-CVIM-231 No.26
2022/11/19

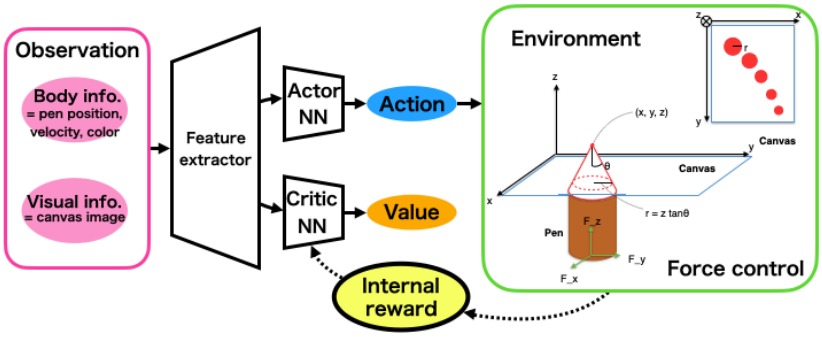

図 1 提案モデルの概要図．力制御によりペンを駆動し，キャンバス上にストロークを描画する環境（右側）の中で，エージェント（左側）が観測を受け取り，行動と価値を出力する．また，内発的動機に基づく報酬の情報を用いて出力を更新する．

究では，アート創作・鑑賞における経験の蓄積によって美の知覚が形成されていく過程に焦点を当て，モデル化・実験・解析を行うものとする．

また，発達研究における構成論的アプローチにおいては，行動を動機づける基本原理を想定し，身体を持ったエージェントに埋め込むと先に述べた．美術は世界についての知識獲得を目的とした視覚脳の機能の延長として存在するという主張 [7] やアートの創作や鑑賞は世界を理解しようとする好奇心に駆動された振る舞いの結果であるという主張 [8] に基づき，本研究では，アート創作・鑑賞を駆動する基本原理として世界に関する知識獲得の内発的動機を想定する．また，身体性はシステムと環境の相互作用において状態と行動が従うべき制約として機能し [5]，システムにとっての情報の入出力関係を規定しうるものであり，それゆえに獲得される振る舞いの差異にも影響を与える [9] ものである．このように，構成論的アプローチの下，内発的動機と身体性を考えることで，何を目的とし，どのような情報を受け取り，どのように外部に働きかけるかによって，エージェントの美の知覚がどのように変化するかを調べることができる．このような視点は画像データに対する人間の美的評価を教師あり手法で学習して，機械が美的評価を行うという枠組みでは扱えていない問題である．

以上を踏まえ，本研究では身体性と内発的動機の影響の下，描画行動を学習する強化学習エージェントを用いたモデルを提案する．強化学習の採用は，人間の美的感覚が報酬系の活動と密接に関連するという神経美学の知見 [10], [11] や，創造的な過程は，試行錯誤の探索プロセスであるという認知科学の知見 [12], [13] を見ても適切であると考えられる．そして，エージェントの美的評価処理の振る舞いは，深層強化学習における価値関数ネットワークの部分に現れるとして，実験・解析を行っていく．価値関数は観測を受け取り，その価値を返す機能を持つものであり，これが人間が刺激を受け取り，それに対して美的評価を下すプロセスに対応すると考えられるからである．ただし，美的評価プロセスとしての価値関数応答の解析に関しては今後の課題とし，本稿では，提案した強化学習エージェントの描画行動が身体性や内発的動機付けにおける設定の相違によってどのように変化するかに関して，調査結果を報告する．

2. 手法

提案モデルの構造は図 1 のようになっている．図 1 右側に示すように，環境は仮想的なキャンバスとペンで構成されており，エージェントが力を加えてペンを動かし，ペンがキャンバスに触れたところに，ペンの z 座標に比例した半径の円が描かれるようになっている．これを短い時間間隔で更新することで，任意の形状のなめらかなストロークを描画することができる．

そして，図 1 左側に示すように，エージェントはペンの位置・速度・色と，現在のキャンバス画像を観測として受け取り，Proximal Policy Optimization アルゴリズム [14] に従って，状態価値と採るべき行動を出力する．状態価値はエージェントの方策更新に利用される一方で，行動は環境に渡され，環境のダイナミクスが 1 時間ステップ進むという仕組みになっている．そして，環境から報酬が返され，エージェントの状態価値推定の改善に利用される．

観測を構成するキャンバス画像は，幅 224 ピクセル，高さ 224 ピクセル，チャネル数 3（BGR）の画像データ，ペンの位置と速度は環境中の x,y,z の 3 つの軸における値のベクトルデータ，ペンの色は BGR 各チャネルのピクセル値 (0 から 255) のベクトルデータである．描画中のストロークの色を一定に保つために，ペンがキャンバスに触れた瞬間から次にペンがキャンバスから離れるまでペンの色が固定されるようになっている．行動は，ペンに加えられる力（x,y,z 方向）と色（BGR）に対応する −1 から 1 の範囲の値で構成される要素数 6 のベクトル形式のデータである．環境はこの行動ベクトルを受け取り，運動方程式 $a = F/m$ に従って微小時間 δ_t の間，ペンに加速度が加えられた後の速度と位置を求め，描画を行う．（ただし，a は加速度，F は加えられた力，m はペンの質量である．）エージェントにとっての 1 ステップはこの微小時間を複数回 n_{split} 繰り返

ⓒ 2022 Information Processing Society of Japan

Vol.2022-CG-188 No.26
Vol.2022-DCC-32 No.26
Vol.2022-CVIM-231 No.26
2022/11/19

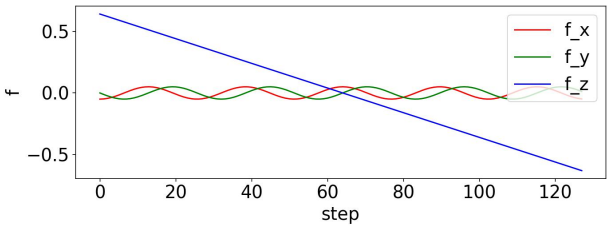

図 2 制御入力（x,y,z 方向の力）パターン

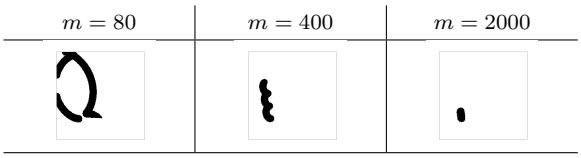

図 3 ペンの質量 $m = 80, 400, 2000$ における描画結果

したものとする．また，微小時間 δ_t の間に，ペンの x,y 座標位置がキャンバスの端 ($x = 0, 224$ や $y = 0, 224$) に到達したとき，ペンの速度と加速度は 0 となるものとしている．ペンの z 座標は $z = 0$ から $z = Z_{\max}$ まで取ることができるとし，$z = 0$ のときはキャンバスに描画は行わないものとしている．例えば，$\delta_t = 5.0$，$n_{\text{split}} = 10$，$Z_{\max} = 10$[pixel] の下，ペンの質量を $m = 80, 400, 2000$ と変え，同じ制御入力（図 2 のような x,y,z 方向の力 $F_{\text{x}}, F_{\text{y}}, F_{\text{z}}$）を 100 ステップ間入力したときの描画結果は図 3 のようになる．ペンの質量の違いで描画結果が大きく変わることがわかる．以下の実験では，$\delta_t = 5.0$，$n_{\text{split}} = 10$，$m = 400$ を基本設定とし，$Z_{\max}$ に関しては，実験 1 では 25[pixel]，実験 2 では 10[pixel] とした．

3. 身体性・内発的動機と描画行動の関連の調査

3.1 実験 1：物理パラメータが描画学習に与える影響

本節では，描画行動を学習する強化学習エージェントを用意し，物理パラメータが描画行動の学習に与える影響を調査する．特に，環境中のペンとキャンバスの間に生じる粘性抵抗の係数に着目する．$f_{\text{vr}} = k_{\text{vr}} v$ に従って生じる粘性抵抗がある状況において，描画行動を学習するエージェントを訓練する．（ただし，f_{vr} は粘性抵抗力，k_{vr} は粘性係数，v はペンの速さである．）エージェントには，各時間ステップ t において，どれだけキャンバス画像に変化を生じさせることができたか（ピクセル値で計算する距離 $d(I_{t+1}, I_t)$）に比例して報酬 r_{dif} を与えることにする．具体的には，$r_{\text{dif}} = c_{\text{dif}} d(I_{t+1}, I_t)$ で計算される．（ただし，c_{dif} は報酬の大きさを制御する定数である．）

3.1.1 具体的な実験設定

ここでエージェントの学習における設定について，具体的に説明する．価値関数ネットワークと方策関数ネットワークはともに共通の特徴抽出層をもっており，そのあとに構造が分岐して，それぞれ価値と行動を出力する．価値推定・方策更新は PPO アルゴリズムに基づいて行い，学習率は 0.0003，割引率 γ は 0.99，価値推定のための係数 λ は 0.95，方策の比のクリッピングの閾値 ϵ は 0.2，勾配のクリッピングの閾値は 0.5 とし，価値推定のためのロールアウトの長さは 1024 ステップとする．これらの設定の下，1 エピソードは 128 ステップからなるものとし，学習は全 5000 エピソード，計 640000 ステップ行う．エピソードの開始時刻に，キャンバスは白紙に戻され，ペンの x,y 座標はキャンバス内のランダムな位置に決定され，z 座標は 0 として設定される．以下の実験においては，5 種類の乱数シードの平均値を用いて解析を行った．

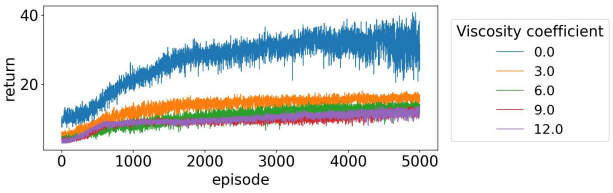

図 4 収益の変化（異なる粘性係数による差異）

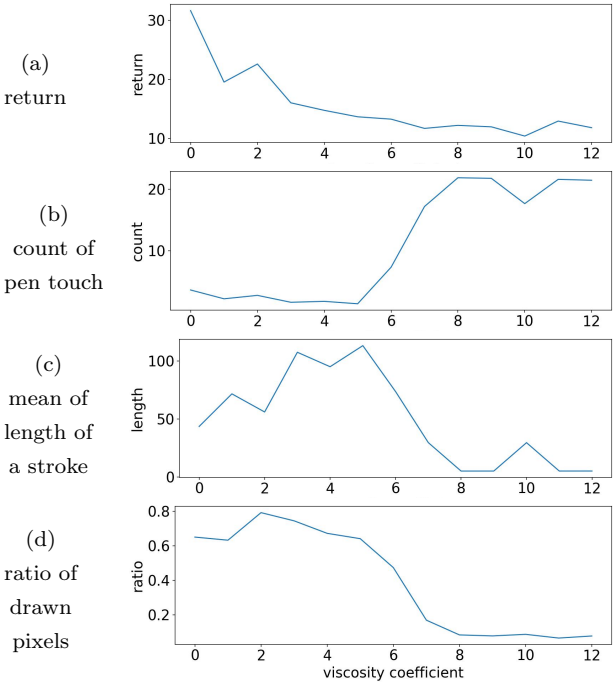

図 5 最後 100 エピソードにおける (a) 収益，(b) ペンの接地回数，(c) ストロークの平均描画時間，(d) 描画面積比の平均値

3.1.2 実験結果と考察

収益の変化は図 4 のようになった．また，学習最終段階（全 5000 エピソード中，最後の 100 エピソード）における収益の平均値を，$k_{\text{vr}} = 0.0$ から 12.0 の範囲で 1.0 ずつ変えた場合に関してプロットすると，図 5(a) のようになった．粘性抵抗が小さくなるほど，ペンを動かすために必要な力の大きさは小さくなるため，描画行動自体が容易となり，その分収益も高くなることが確認できた．

また，描画行動がどのように変化するかを調べるため，描画行動の指標として，1 エピソード内におけるペンの接地回数，ストロークの平均描画時間，描画表現の指標とし

ⓒ 2022 Information Processing Society of Japan

Vol.2022-CG-188 No.26
Vol.2022-DCC-32 No.26
Vol.2022-CVIM-231 No.26
2022/11/19

て，1 エピソード終了時刻における描画面積比，という 3 つの指標を提示する．ペンの接地回数は，1 エピソード内において，ペンがキャンバスに接地した（z 座標が 0 より大きくなった）回数をカウントしたもの，ストロークの平均描画時間は，1 エピソード内において，ペンがキャンバスに接地している期間のステップ長の平均値，描画面積比は，1 エピソードの終了時刻におけるキャンバス画像と，開始時刻のキャンバス画像との差分から，キャンバス全体に対してストロークが描画された領域の面積はどれくらいかを計算したものである．$k_{\mathrm{vr}} = 0.0$ から 12.0 の範囲で 1.0 ずつ変えた場合に関して，学習最終段階（全 5000 エピソード中，最後の 100 エピソード）におけるペンの接地回数，ストロークの平均描画時間，描画面積比の平均値はそれぞれ，図 5(b)，(c)，(d) のようになった．粘性係数が 5.0 以下の場合は接地回数が 5 回以下，ストロークの平均描画時間は 40 以上あり，描画面積比も 0.6 以上あることが分かる．一方で，粘性係数が 6.0 以上の場合は接地回数が上昇し，ストロークの平均描画時間が短くなり，描画面積比も小さくなることが分かる．

各粘性係数の場合の描画結果は図 6 のようになった．この図は，各粘性係数の場合に，全 640000 ステップの訓練の途中におけるエージェントモデルを取り出して，描画をさせたときの結果である（この図は 1 シードの実行結果を取り出したものであり，実際はシードごとにばらつきがある）．$k_{\mathrm{vr}} = 3.0$ のときは，学習が進むにつれて，キャンバスに最も大きい変化を加えられる黒いストロークを全体に描くようになったことが分かる．また，$k_{\mathrm{vr}} = 0.0$ のときは，学習序盤からすでに黒いストロークを全体に塗り付ける挙動が獲得できており，その後，黒いキャンバスの上にさらに他の色を塗り重ねて，追加の報酬を獲得するところまで学習が進んでいったことが読み取れる．最終的には，最も効率的にキャンバスのピクセル値の変化を加えられる白いストロークを塗り重ねる挙動が獲得されている．

ストロークの平均描画時間（図 5(c)）において，$k_{\mathrm{vr}} = 3.0, 4.0, 5.0$ の場合に比べて $k_{\mathrm{vr}} = 0.0, 1.0, 2.0$ の場合の方がストロークの平均描画時間が短い理由は，上の場合のように $k_{\mathrm{vr}} = 0.0, 1.0, 2.0$ のとき，黒いストロークの描画の後，ストロークの色を変えるためだと考えられる．

また，図 6 を見ると，$k_{\mathrm{vr}} = 6.0$ のときは，短い点状のストロークを色を変えながら描き重ねる挙動が生じ，$k_{\mathrm{vr}} = 12.0$ のときはほぼすべてのストロークが点状のものになっていることが分かる．描画面積比（図 5(d)）において，$k_{\mathrm{vr}} = 6.0$ 以上のとき，描画面積比が大きく減少しているのは，粘性抵抗が大きい状況下で，ペンを x,y 軸方向に移動させて描画をするよりも，ペンを上下して，色を変えながら塗り重ねてキャンバスに変化を加えたほうが，高い報酬を獲得できることを学習したためと考えられる．

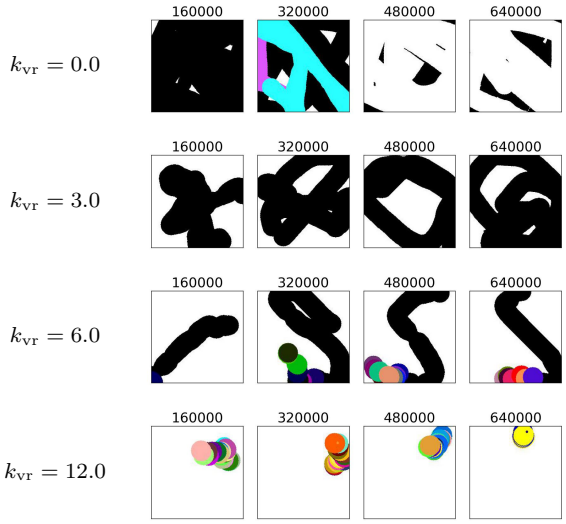

図 6 描画結果の変化（異なる粘性係数による差異）．画像上の数字は経過した学習ステップ数を表す．

3.1.3 実験 1 のまとめ

本実験 1 を通して，粘性係数が小さいほどストロークを描くことが容易なため，より大きな報酬をもらえる描画行動（長く黒いストロークの描画，白いストロークの塗り重ね）が獲得されるということ，また，粘性係数が一定の値より大きくなると描画行動が大きく変化し，長いストロークを描くよりも短いストロークで色の塗り重ねによって報酬を獲得しようとするように戦略が変化することが分かった．力制御によって描画を行う環境における強化学習エージェントの描画行動の学習（難易度や獲得される振る舞い）は，粘性抵抗の大きさという物理的設定に依存しているということができる．また，ペンの接地回数・ストロークの平均描画時間，描画面積比は，力制御によって描画を行う環境の分析用指標として有用であることも確認できた．

3.2 実験 2：内発的動機付けの影響

本節では，内発的動機付けを用いてエージェントに描画行動を学習させたときの振る舞いを調査する．Aubret et al. によれば，内発的動機付けの定式化手法は様々あるが，知識獲得を目的とした手法とスキル学習を目的とした手法の 2 つに大別されるという [15]．このうちの前者は，エージェントが環境に関する新しい知識（操作可能なものは何か，世界がどのように動くのかという情報）を獲得することを動機とする定式化手法である．この手法の中でも有名なものとして，Intrinsic Curiosity Module (ICM) [16] や Random Network Distillation (RND) [17] がある．

ICM は，状態から特徴を抽出する特徴抽出器と，現在時刻に採られた行動を予測する逆モデル，次時刻の状態を予測する順モデルからなる．はじめに，現在時刻 t の状態 s_t と次時刻 $t+1$ の状態 s_{t+1} をそれぞれ特徴抽出器に入力し，中間表現 $\phi(s_t)$ と $\phi(s_{t+1})$ を得る．この二つを逆モデ

ⓒ 2022 Information Processing Society of Japan

Vol.2022-CG-188 No.26
Vol.2022-DCC-32 No.26
Vol.2022-CVIM-231 No.26
2022/11/19

ルに入力し，現在時刻において採られた行動 a_t の予測 $\hat{a}_t$ を得る．次に，現在時刻の状態に対する中間表現 $\phi(s_t)$ と行動 a_t を順モデルに入力し，次時刻の状態に対する中間表現 $\phi(s_{t+1})$ の予測 $\hat{\phi}(s_{t+1})$ を得る．順モデルは $\phi(s_{t+1})$ と $\hat{\phi}(s_{t+1})$ を近づけるように学習させ，逆モデルは a_t と $\hat{a}_t$ を近づけるように学習させる．このような設定の下，各時刻 t における内発的報酬は，$\phi(s_{t+1})$ と $\hat{\phi}(s_{t+1})$ の差分として定義される．ICM は環境のダイナミクスに関してどれほどの知識を持っているかを予測誤差を用いて定量化し，内発的報酬として利用する手法である．

RND はニューラルネットワークの蒸留という手法を用いて内発的報酬を計算する手法である [17]．はじめに，構造は同一であり，かつ，それぞれランダムな重みで初期化されたニューラルネットワークを 2 つ用意する．一方（target ネットワークと呼ぶ）の重みは常に固定しておき，もう一方（predictor ネットワークと呼ぶ）の重みを学習により最適化する．最適化の目標は，ある時刻の環境の状態を target ネットワークと predictor ネットワークの双方に入力した時に，両者の出力を近づけるようにすることである．両者の出力の誤差を内発的報酬として利用する．この手法は，一度見たことのある状態に対する target ネットワークと predictor ネットワークの出力は，全く新しい状態に対する両者の出力よりも誤差が小さくなるという仮説から着想を得たものである．RND は観測の新規性を定量化する手法であり，新規な観測を目指した探索は，環境についての知識獲得を動機としたものとみなすことができる．

3.2.1 具体的な実験設定

本実験では，知識獲得の内発的動機付け手法として ICM や RND を利用する．以下では，環境が部分観測であることを前提とし，状態の代わりに観測を用いて説明する．

ICM を構成する特徴抽出器，順モデル，逆モデルの学習が一定のエピソード間隔 T_{interval} ごとに行われる．全体の損失は $l_{\text{total}} = c_{\text{fm}} l_{\text{fm}} + c_{\text{im}} l_{\text{im}}$ と表される．（ただし，l_{fm} は順モデルの予測誤差，l_{im} は逆モデルの予測誤差であり，c_{fm} と c_{im} はそれらの比率を決める係数である．）実験では，$(c_{\text{fm}}, c_{\text{im}}) = (0.5, 0.5)$ とした．学習においては，逆モデルの損失による勾配は特徴抽出器まで流す一方で，順モデルの損失による勾配は順モデル内のみで流し，特徴抽出器までは流さない．ICM は観測と行動から次時刻の観測の特徴量表現を予測する手法であり，できるだけ観測における情報が多い方が望ましいので，キャンバス画像・ペンの位置・速度・色のすべての情報を ICM に入力するものとする．

RND においても学習は一定のエピソード間隔 T_{interval} ごとに行われる．RND は，観測自体の新規性を評価する手法である．したがって，RND の入力に観測情報のすべてを利用してしまうと，キャンバス画像以外の要素（ペンの位置・速度・色）における新規性を求めてしまい，キャンバスにおける新しいストロークの描画が生じにくくなる（予備実験によってその傾向が見られた）．したがって，以下の実験では，RND に入力する観測情報の要素はキャンバス画像に限定するものとする．ただし，強化学習エージェントにとっての観測情報の要素は変わらずに，キャンバス画像・ペンの位置・速度・色であることに注意されたい．

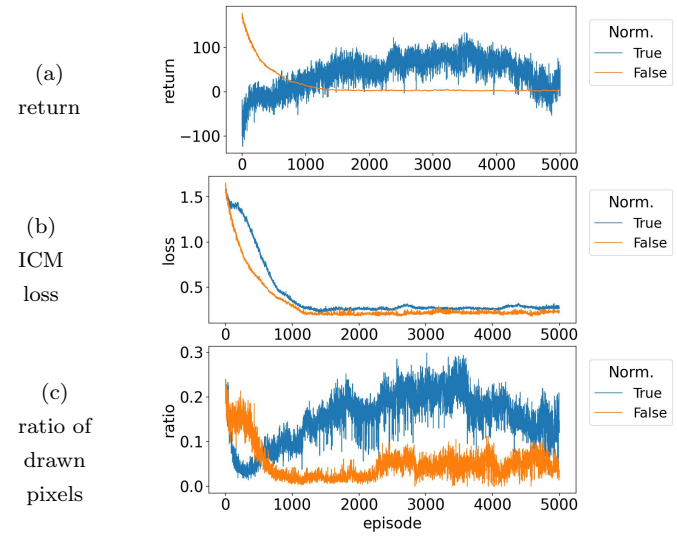

図 7 (a) 収益，(b)ICM の損失，(c) 描画面積比の平均値，の変化（ICM 報酬正規化の有無による差異）

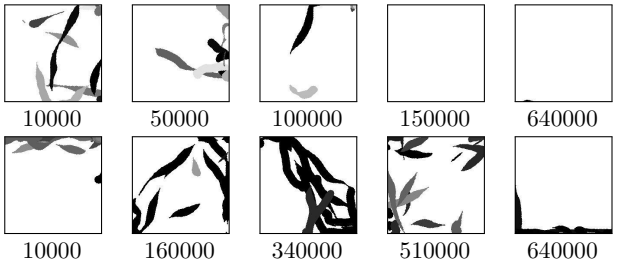

図 8 描画結果の変化（ICM 報酬正規化の有無による差異）．数字は経過した学習ステップ数を表す．上段が報酬正規化なしの場合，下段が報酬正規化ありの場合．

実験では，内発的動機付け手法として ICM，RND を用い，エピソード内での報酬の平均と標準偏差を用いて報酬の正規化を行うか否かの影響や，学習のエピソード間隔の差異 $T_{\text{interval}} = 1, 10, 100$ の影響を調査した．

その他エージェントの学習に関するハイパーパラメータ設定（学習率やエピソード長，訓練ステップ数等）は，実験 1 と同じもの（小節 3.1.1 参照）を使用した．また，解析の簡単化のため，ペンの色はグレースケールに固定した．解析は 5 種類の乱数シードの平均値を用いて行った．

3.2.2 実験結果と考察

＜ICM による内発的報酬の正規化の有無＞

はじめに，学習のエピソード間隔 $T_{\text{interval}} = 1$ の下，ICM による内発的報酬の正規化の有無の影響について調査した．収益，順モデル・逆モデルの予測誤差による合計損失のグラフはそれぞれ図 7(a)，(b) のようになった．内発的報酬

ⓒ 2022 Information Processing Society of Japan

Vol.2022-CG-188 No.26
Vol.2022-DCC-32 No.26
Vol.2022-CVIM-231 No.26
2022/11/19

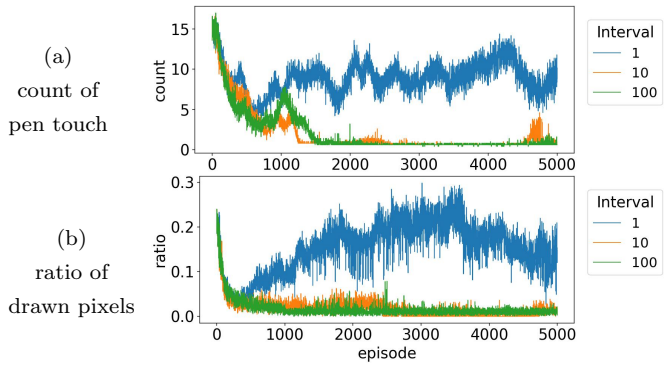

図 9 (a) ペンの接地回数，(b) 描画面積比，の変化（ICM 学習のエピソード間隔による差異）

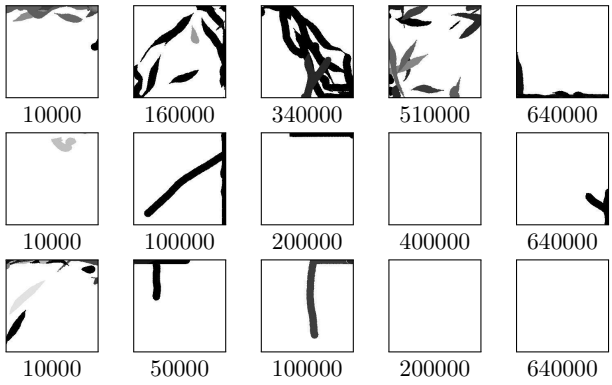

図 10 描画結果の変化（ICM 学習のエピソード間隔による差異）．画像下の数字は経過した学習ステップ数を表す．上段が $T_{\text{interval}} = 1$，中段が 10，下段が 100 に対応する．

の正規化が無い場合は，最初の 1000 エピソードでの更新によって，予測誤差の合計損失と内発的報酬が大きく減少し，0 に近い値で収束してしまっている様子が確認できる．この場合の描画表現の変化は図 7(c)，図 8 上段のようになった．シードによってばらつきはあるものの，多くのシードでこのような描画しない方向に学習が進んでいた．一方，内発的報酬の正規化を行う場合は，収益が 0 に収束することはなく，描画表現（図 7(c)，図 8 下段）を見ても，描画行動が継続して現れることが確認できる．また，途中で描画スタイルの変遷が生じているようにも見える．例えば，学習 10000 ステップ（すなわち $10000/128 \approx 78$ エピソード）における描画行動は，キャンバスの端に短いストロークを重ね描くような挙動であるが，学習 160000 ステップ（1250 エピソード）における描画行動は，キャンバスの中央部分にも短いストロークを描き始めている．そして，学習 340000 ステップ（約 2656 エピソード）においては，長いストロークを用いた表現になっており，学習 510000 ステップ（約 3984 エピソード）においては再度短いストロークを用いた表現に戻り，そして学習 640000 ステップ（5000 エピソード）においては，キャンバスの端にストロークを重ねる挙動に戻っている．ICM という環境ダイナミクスに関する新しい知識獲得を目指した手法によって，力制御環境による描画を行うエージェントにおいても，学習の中で様々な描画行動が発現しうることが確認できた．

＜ICM モジュールの訓練間隔＞

ICM による内発的報酬の正規化のありの設定の下，学習のエピソード間隔 $T_{\text{interval}} = 1, 10, 100$ の影響について調査した．ペンの接地回数，描画面積比のグラフはそれぞれ図 9(a)，(b) のようになった．これらを見ると，学習のエピソード間隔が大きくなると，ペンの接地回数・描画面積比が 0 付近に収束することがわかる．実際にあるシードにおける描画行動（図 10）を見てみると，$T_{\text{interval}} = 10, 100$ の場合は，最初の 100000 ステップあたりの時点でキャンバスの端にストロークを持って行くのみの描画行動になっており，その後学習が進んでも，キャンバス中央部にストロークを描く挙動が現れることはほぼなかった．

＜RND による内発的報酬の正規化の有無＞

学習のエピソード間隔 $T_{\text{interval}} = 1$ の下，RND による内発的報酬の正規化の有無の影響について調査した．収益，RND の損失の変化のグラフはそれぞれ図 11(a)，(b) のようになった．内発的報酬の正規化が無い場合は，初期のエピソードでの更新によって，予測誤差の合計損失と内発的報酬が大きく減少し，ある値に収束してしまっている様子が確認できる．実際に，図 11(c) を見ると，描画面積比はほぼ 0 に近い値に収束しており，描画結果の変化（図 12 上段）を見ると，ほぼ何も描かない挙動に収束している．

一方，内発的報酬の正規化を行う場合は，予測誤差の合計損失は 0 に近づいているが，収益は減少することなく，高い値（100 から 150 程度）で推移している．これは，報酬の正規化を行う場合は，予測誤差の合計損失が 0 に近づいたとしても，エピソード内での予測誤差の平均値に対して，常により大きい予測誤差を出すような振る舞いに正の報酬が与えられるため，RND が予測がわずかにうまくいっていない領域を発見し，そこに変化をもたらすような行動を学習した結果だと考えられる．描画面積比（図 11(c)）は 2000 エピソードあたりまでは上昇し，その後下降に転じている．描画の振る舞い（図 12 下段）は，最初様々な色の短いストロークを描く挙動であったが，途中の 150000 ステップあたりでは黒く長いストロークを一面に描く挙動になっている．そして再度複数の色を用いるような挙動となり，最終的に，短いストロークを描画するようになっている．このように，様々な描画行動が変遷していく様子が確認できる．短いストロークを描く挙動が学習初期だけでなく学習終盤にも現れた理由については，以下のように予想できる．RND を構成するニューラルネットワークは，データのバッチを用いた重みの更新によって学習するため，直近に学習したデータに対する予測は精度が高く，時間的に離れた過去に学習したデータに対する予測は比較的精度が低くなる．したがって，黒く長いストロークの描画が続いた状況では，複数色の短いストロークからなるキャンバス画像に対する予測精度が低くなり，その分報酬が得られる

ⓒ 2022 Information Processing Society of Japan

Vol.2022-CG-188 No.26
Vol.2022-DCC-32 No.26
Vol.2022-CVIM-231 No.26
2022/11/19

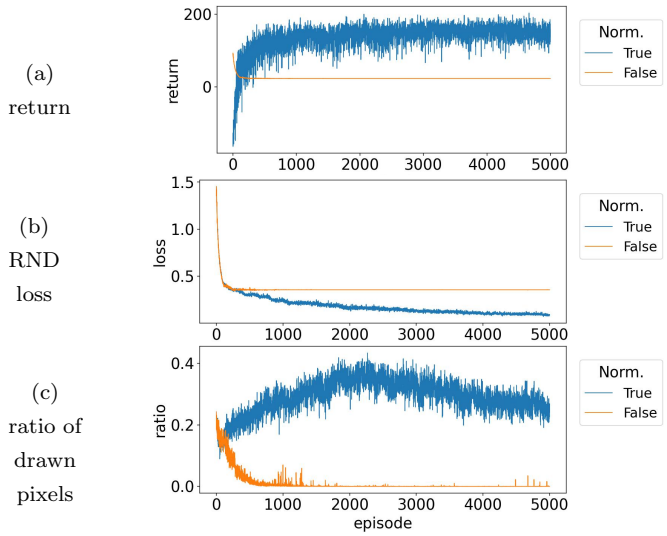

図 11 (a) 収益，(b)RND の損失，(c) 描画面積比の平均値，の変化（RND 報酬正規化の有無による差異）

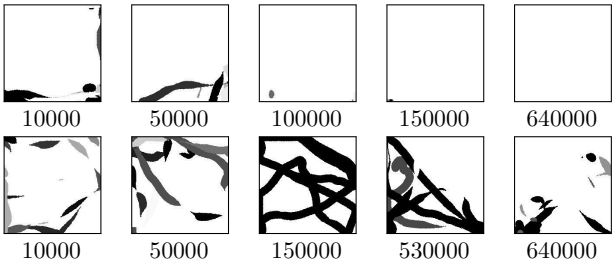

図 12 描画結果の変化（RND 報酬正規化の有無による差異）．数字は経過した学習ステップ数を表す．上段が報酬正規化なしの場合，下段が報酬正規化ありの場合．

ため，次第に複数色の短いストロークからなるキャンバス画像を描画するように遷移したと考えられる．

＜RND モジュールの訓練間隔＞

RND による内発的報酬の正規化ありの設定の下，学習のエピソード間隔 $T_{\text{interval}} = 1, 10, 100$ の影響について調査した．収益，描画面積比のグラフはそれぞれ図 13(a)，図 13(b) のようになった．学習のエピソード間隔が大きくなると，描画面積比が小さくなり，収益も小さくなる傾向があることがわかる．実際のあるシードにおける描画結果（図 14）からは，$T_{\text{interval}} = 1$ の場合にキャンバス上での描画が継続し，描画行動のパターンが変化している様子が見られる．これは，描画面積比のグラフが折れ線状になっている様子からも確認できる（図 13(b)）．一方で $T_{\text{interval}} = 100$ の場合は学習初期から，キャンバスの端でストロークを塗り重ねる挙動に収束していた．

＜RGB カラーのストロークを用いた場合＞

最後に，RGB カラーのストロークを用いて，ICM や RND によって内発的動機付けを行った場合の描画行動を紹介する．いずれの場合も，学習のエピソード間隔 $T_{\text{interval}} = 1$，内発的報酬の正規化ありの場合で学習させた．描画結果は図 15 のようになった．RGB カラーのストロークを用いた

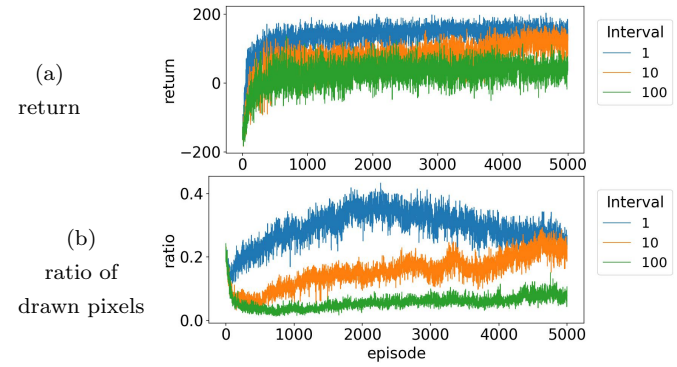

図 13 (a) 収益，(b) 描画面積比，の変化（RND 学習のエピソード間隔による差異）

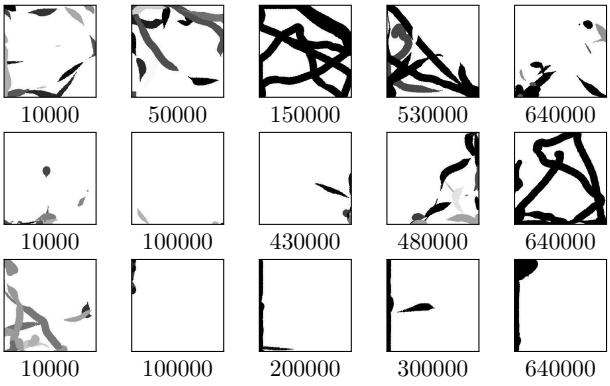

図 14 描画結果の変化（RND 学習のエピソード間隔による差異）．画像下の数字は経過した学習ステップ数を表す．上段が $T_{\text{interval}} = 1$，中段が 10，下段が 100 に対応する．

場合であっても，複数色のストロークの描画行動が継続して獲得されていることが確認できる．

3.2.3 実験 2 のまとめ

本実験 2 では，力制御によりペンを駆動し，キャンバス上に描画を行う環境において，内発的動機づけによって，エージェントのどのような描画行動を発現させられるかを調査した．内発的動機付け手法として ICM と RND を採用し，いずれの場合も，エピソード内の平均と標準偏差を用いて報酬を正規化し，短いエピソード間隔でモジュールを学習させた方が 1 つのエピソード内で様々な描画行動を生じさせることができると分かった．現状は，内発的動機付けによって駆動された描画行動を調査したにすぎず，美の知覚の形成との関連については調査・考察できていない．この点は今後の課題である．

4. 結論と展望

本稿では，力制御によりペンを駆動し，物理法則に従いながらキャンバス上にストロークを描画するような環境において，環境の物理パラメータが描画行動やその学習にどのような影響を与えるかについて調査した．また，描画行動の原動力として，内発的動機付けという仕組みに着目し，強化学習の分野で採用されている定式化を用いて，経験を通して描画行動を学習するエージェントに組み込み，どの

ⓒ 2022 Information Processing Society of Japan

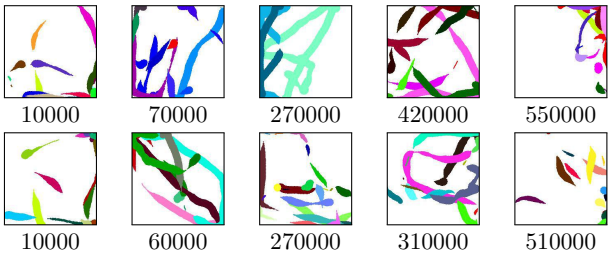

図 15 描画結果の変化（カラーストロークを用いた場合）．画像下の数字は経過した学習ステップ数を表す．上段が ICM の場合，下段が RND の場合に対応する．

ような描画行動・描画表現が発現するかを調査した．これまでに画像生成を行う機械学習・AI モデルは数多く提案されているが，本研究で提案したような身体性と内発的動機を持ったエージェントによる画像生成を行った事例は筆者の知る限りでは存在していない．AI エージェントにアート作品を作らせることができるかに関しては議論があるため（例えば [18]），本稿で提案したモデルの生成した画像がアートであると主張することは避けるが，本稿での提案モデルは AI エージェントによる自律的なアート作品創作に近づくための第一歩としての意義がある．今後はこのモデルを使用し，先述した本研究の目的である，AI エージェントにとっての美的評価基準が学習過程でどのように形成されていくかについて調査を進めていく予定である．

謝辞 本研究の一部は（株）電通との共同研究費「クリエイティブの科学に向けた探索的研究」の支援を受けた．

参考文献

[1] Kao, Y., Wang, C. and Huang, K.: Visual Aesthetic Quality Assessment with A Regression Model, *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 1583–1587 (2015).
[2] Elgammal, A., Liu, B., Elhoseiny, M. and Mazzone, M.: CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms (2017). arXiv:1706.07068.
[3] 徳井 直生：創るための AI　機械と創造性のはてしない物語，ビー・エヌ・エヌ (2021).
[4] 石津 智大：神経美学ー美と芸術の脳科学, 共立出版 (2019).
[5] Kuniyoshi, Y.: Fusing autonomy and sociality via embodied emergence and development of behaviour and cognition from fetal period, *Philosophical Transactions of the Royal Society B: Biological Sciences*, Vol. 374, No. 1771, p. 20180031 (2019).
[6] Leder, H., Belke, B., Oeberst, A. and Augustin, D.: A model of aesthetic appreciation and aesthetic judgments, *British Journal of Psychology*, Vol. 95, No. 4, pp. 489–508 (2004).
[7] セミール ゼキ：脳は美をいかに感じるかーピカソやモネが見た世界，日本経済新聞社 (2002). 河内十郎 監訳.
[8] Schmidhuber, J.: Simple Algorithmic Theory of Subjective Beauty, Novelty, Surprise, Interestingness, Attention, Curiosity, Creativity, Art, Science, Music, Jokes, 計測と制御，Vol. 48, No. 1, pp. 21–32 (2009).
[9] 國吉 康夫，寒川 新司，塚原 祐樹，鈴木 真介，森 裕紀：人間的身体性に基づく知能の発生原理解明への構成論的アプローチ，日本ロボット学会誌，Vol. 28, No. 4, pp. 415–434 (2010).
[10] Ishizu, T. and Zeki, S.: Toward A Brain-Based Theory of Beauty, *PLOS ONE*, Vol. 6, No. 7, pp. 1–10 (2011).
[11] Levy, D. and Glimcher, P.: Common Value Representation—A Neuroeconomic Perspective, *Handbook of Value: Perspectives from Economics, Neuroscience, Philosophy, Psychology and Sociology* (Brosch, T. and Sander, D., eds.), Oxford University Press, pp. 85–118 (2015).
[12] 阿部 慶賀：創造性はどこからくるか：潜在処理、外的資源、身体性から考える，2, 共立出版 (2019).
[13] 開 一夫，鈴木 宏昭：表象変化の動的緩和理論：洞察メカニズムの解明に向けて，認知科学，Vol. 5, No. 2, pp. 69–79 (1998).
[14] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O.: Proximal Policy Optimization Algorithms (2017). arXiv:1707.06347.
[15] Aubret, A., Matignon, L. and Hassas, S.: A survey on intrinsic motivation in reinforcement learning (2019). arXiv:1908.06976.
[16] Pathak, D., Agrawal, P., Efros, A. A. and Darrell, T.: Curiosity-driven Exploration by Self-supervised Prediction, *Proceedings of the 34th International Conference on Machine Learning* (Precup, D. and Teh, Y. W., eds.), Proceedings of Machine Learning Research, Vol. 70, PMLR, pp. 2778–2787 (2017).
[17] Burda, Y., Edwards, H., Storkey, A. and Klimov, O.: Exploration by Random Network Distillation (2018). arXiv:1810.12894.
[18] Hertzmann, A.: Can Computers Create Art?, *Arts*, Vol. 7, No. 2 (2018).

ⓒ 2022 Information Processing Society of Japan