展　望

ロボティクスと強化学習

Robotics and Reinforcement Learning

浅　田　　稔*1*2　*1大阪国際工科専門職大学　*2大阪大学先導的学際研究機構

Minoru Asada*1*2

*1International Professional University of Technology in Osaka　*2Open and Transdisciplinary Research Initiatives, Osaka Univeristy

1. はじめに

強化学習はエージェントが環境と相互作用を通じて，所期のタスクを達成する行動学習法 [1] であり，エージェントとしてロボットが想定され，実世界のロボットの行動学習法として利用されてきた．筆者らは，20 年以上前，ロボカップの当初から，強化学習法をロボカップのタスクドメインに適用し，多くの論文を発表してきた（例えば，書籍 [2] の 2～5 章，7～10 章）．それらは，通常の強化学習の適用から発展して，状態行動空間の構成，環境複雑度制御，階層構造，マルチエージェントへの拡張など，強化学習のみならず，ロボットの行動学習一般に関わる課題を追求してきた．これらを通じて，実環境におけるロボットの学習の適用と課題についても触れている [3]．すでに古典的な成果ではあるが，ロボットの実環境への適用に関しては，言語表現がことなるものの，本質的にはさして変わっていないと察する．この間の一番大きな衝撃は，深層学習の興隆により，強化学習にも多大な影響を与えた点である．すなわち，当初の課題であった状態行動空間の構成が状態表現課題に，模倣学習が逆強化学習の枠組みに，環境構造の制御が事前知識の取り扱いなどに組み込まれ，実環境でのロボット学習の困難さとその対応に表われている．以下では，関連論文約 250 編のレビュー論文 [4]，最近の話題の解説 [5]，また，実環境でのロボット強化学習で実績の多い，University of California, Berkeley の成果から最近の論文 [6] を参考に強化学習とロボティクスを展望する．最後に，より高度な認知機能への進展の可能性について論じる．

2. 強化学習の機械学習としての位置づけ

書籍 [7] の第 4 章では，教師あり学習，教師なし学習の間に，環境が教師の役割を果たす学習法として，強化学習が位置づけされ，脳部位との関連では，小脳，大脳，大脳基底核がそれぞれ，教師あり学習，教師なし学習，強化学習に対応するとされている．教師あり学習学習で始まった深層学習ではあるが，これらのカテゴリー分けと関係なく，それぞれを強化すると同時に境界も曖昧にしつつあると考えられる．そのため，Kober et al. [4] は，やや粗っぽい構造化として，環境やタスクの複雑度，ならびに報酬構造の複雑度の観点から，最も困難な問題として強化学習を位置づけている．

最適制御と強化学習の関係でも，前者が完全知識を前提にしてモデル化していることに対し，後者が解析困難なケースをデータ駆動型で近似を使って問題に対処としているが，理論的には，類似性が指摘されている [8]．強化学習のアプローチ（価値関数，方策探索，関数近似）については，書籍 [1] の 1，2 章を参照されたい．

3. ロボティクスから観た強化学習の課題

強化学習の当初の多くの研究は状態行動空間が離散で完全観測可能なグリッドワールドをタスクドメインとしていたが，実環境でタスクを遂行するロボットの環境条件とは著しく異なり，このことが実応用を困難にしていた．それは，

- 高次元連続状態行動空間
- 時間的にも空間的にも部分観測
- ノイズの影響
- 不完全なモデル化とそれによるモデルの不確実性
- 報酬関数の設計

である [4]．これらは，筆者が 20 年以上前に指摘した課題 [3] とあまり変わっていないが，その対処法は大きく進展している．それらを見ていこう．

3.1 次元の呪い

先にも示したように完全観測のグリッドワールドでは，状態数が数 10 のオーダーでエージェントの上下左右の移動の 4 自由度程度である．これに対し，実世界で作動するロボットとして，空気圧アクチュエータで動くヒューマノイドプラットフォームの CB2（図 1）の場合，ステレオカメラの画像，ステレオマイクロフォンの聴覚，全身およそ 200 個の触覚センサを有し，全身 56 個のアクチュエータで駆動する．いずれも高次元，連続量であり，生データの状態では，巨大なデータ量で，なんらかの工夫が必要である．まさしく次元の呪いである．

3.2 実世界からのサンプル取得

実時間で作動し，時間遅れを考慮した制御アルゴリズムが必要である．加えて，ロボットのメンテナンスにはコストがかかる．また，学習の実験を実施するうえで，初期位置に戻すなどの実験者の介入が必要である．さらに，実験

原稿受付　2021 年 6 月 9 日
キーワード：Model-free, Model-based, Value Function-based, Policy Search Methods, Real World Complexity
*1〒530–0001 大阪市北区梅田 3–3–1
*2〒565–0871 吹田市山田丘 1–1
*1Kita–ku Osaka–shi, Osaka
*2Suita–shi, Osaka

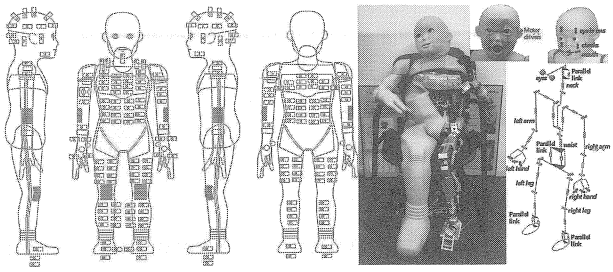

図 1 CB2 の感覚と運動の自由度 [9] [10]

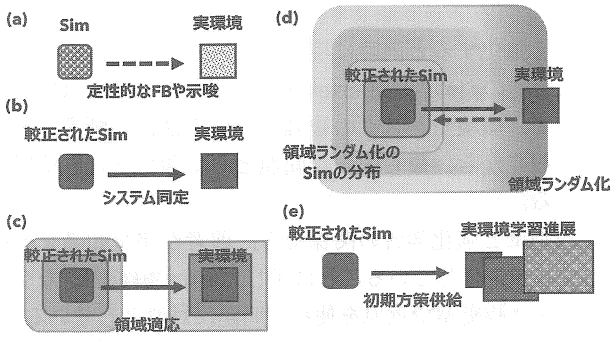

図 2 シミュレーションから実環境への転移の構造（Open AI の Lilian Weng 氏による Web 上の説明図 Fig. 1 から編集拡張）

環境が動的に変化し，ロボット自身のダイナミックな動きを静止することもできない．よって，所望の行動を再現することも困難である．これらに加え，ノイズの影響も多く，全体のコスト（時間，役務，財政）が高く，研究者にとってハイリスクとなる．

3.3 不完全なモデル化とそれによるモデルの不確実性

この課題は，解説 [3] で指摘した問題である．上記の二つの課題から，実環境の実験の前にシミュレーションであたりを付けることがよく行われている．図 2 (a) が，これに対応する．この課題でも深層学習の興隆によるシミュレーション自体の改善がなされている．Open AI の Lilian Weng 氏による Web 上の説明†では，システム同定，領域適応，領域ランダム化の三つのカテゴリーに分けており，同図 (b–d) にそれを示す．図からも分かるように，領域適応は，システム同定を含み，領域ランダム化は領域適応を含んでいる．解説 [3] の説明の用語では，いずれも直接移植型に対応しつつも，シミュレーションの精度が，歴然として上がっている．以前は，設計者によるマニュアルチューニングがほとんどで，論文に明記されないことも多々あったが，近年では，再現性の観点から，明記されると同時に自動チューニングの方向に移行しつつある．その典型が，ランダム化と称されているものである．これは，実験環境の様々なパラメータをランダムに振ることで，実環境の変動に耐えうるシミュレーション結果を導き出すものである．彼女が引用している一つに，実データとの照合による最適化が Chebotar, et al. [11] によって行われている．彼らは，均一なランダム化のシミュレーションで求めた初期方策によって，シミュレーションと実環境でロボットを駆動し，その軌跡の差を最小化するためのランダム化によるシミュレーションの修正を繰り返し，ロボット自身のパラメータ変動にも耐えうる汎化性能を謳っている．引き出し作業やペグの振り入れ作業に適用した結果がビデオで示されている††．

シミュレーション結果を初期方策として用いることは同じだが，実環境でそれを発展させるアプローチもある [12] (同図 (e))．これは，パッサーとシューターの協調行動のマルチエージェント強化学習で，時間を要する下位の行動学習および初級から中級程度へのスキル向上をシミュレーションで実施し，それを実環境に持ち込んで，協調行動の上級スキルを獲得している．解説 [3] の説明の用語では，有機的結合とよんでいる．また，精緻なシミュレーションを実施しつつも，あくまで実環境のみでの学習にこだわるアプローチ [6] もあり，5 章で紹介する．厳密には，シミュレーションと実環境の乖離型であるが，高度な意味合いでの同図 (a) に対応すると言えるだろう．

3.4 ゴールの指定

一般に軽視されがちな課題だが，通常，実験者が監視することで，ゴール状態をロボットに指定している場合が多いが，この課題は，実は報酬関数をどのように設計するかという重要な課題である．初期の実験では，吸収ゴール（absorbing goal）として，ゴール状態のみに価値 1 を割当，学習過程を通じて価値を求めることが多いが，状態すべてを網羅することは時間のコストが高く，適切な報酬関数を定めることが肝要である．サッカーロボットの学習シミュレーション [13] でゴールに到達する途中に 0.5 程度の報酬を設置すると，そこにとどまり出てこない状況，すなわち極大値にトラップされた経験があり，この課題を痛感した．

本稿では，以下でも述べるように，報酬関数の推定に関連して，逆強化学習の手法を用いることで，この課題の解法への取り組みを紹介する．

4. 実環境強化学習実現への様々な工夫

前章で指摘した実環境強化学習実現への課題をなんらかの形で解決する手法の味噌は，効率的表現，ほぼ正確なモデル，そして事前知識や情報の利用である [4]．最初の課題は，古くは状態行動空間の構成課題に対応し，最近では，状態表現学習（State Representation Learning）とよばれている [5]．ここに深層学習の影響が大きい．画像そのものを入力可能にすることで，実環境における多次元情報の取り扱いが容易になった．強化学習過程における価値関数の推定において，離散的なテーブル参照ではスケールアップできず，関数近似手法が多用されている．ここでは価値関数近似を紹介する．方策は学習過程に大きな影響を及ぼすが，ここに事前知識や情報を組み込むこと，すなわち事前に構造化された学習方策が有用性を示す．以下では，これら三つについて外観する．

4.1 状態行動空間構成

状態行動空間の量子化に関する各種の手法については，

†https://lilianweng.github.io/lil-log/2019/05/05/domain-randomization.html
††https://sites.google.com/view/simopt

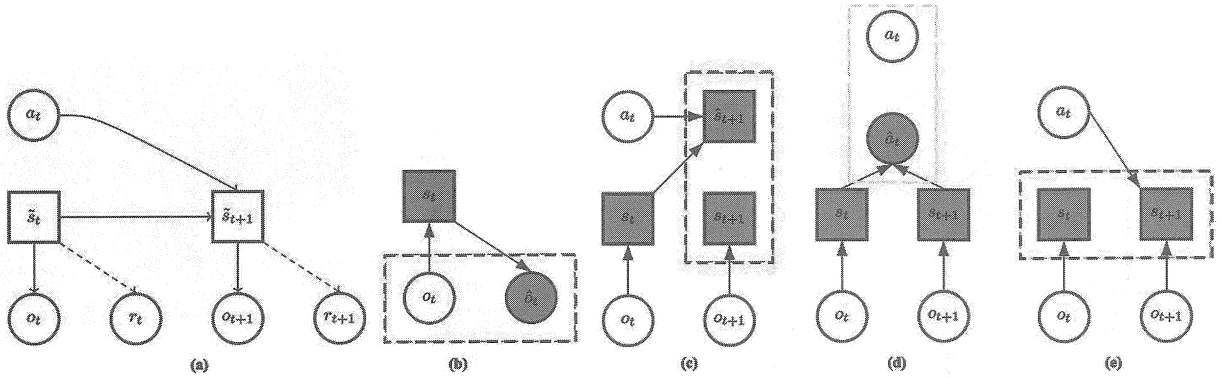

図 3　状態表現学習モデル：(a) 一般グラフィカルモデル，(b) オートエンコーダ，(c) 順モデル，(d) 逆モデル，(e) 事前知識付きモデル（文献 [14] の Figures 1–5）

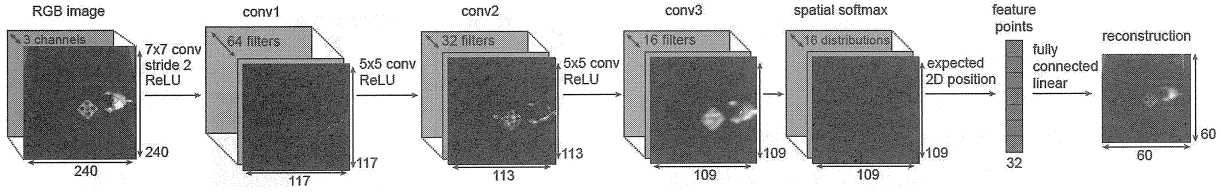

図 4　深層空間オートエンコーダ（文献 [15] の Fig. 3）

文献 [4] の 4.1 節に詳しい．当初，多くの手法では，設計者が学習可能な空間を手作りで構成していたが，実環境では，マルコフ過程を満足する状態行動空間を構成することは難しく，そのため，様々な工夫がなされてきた．例えば，ロボカップの移動ロボットでは，知覚空間におけるボールやゴールの大小の変化が必ずしも一つの運動指令（運動プリミティブ）に対応しないので，状態が変化するまで同一のプリミティブを繰り返し，その一連の動きを一つの運動指令とみなしたり（書籍 [2] 第 2 章），ゴール状態（もしくはすでに状態として切り出されたもの）に同一のプリミティブで到達可能な知覚状況の集合を一つの状態として切り出し，プリミティブの行動系列を一つの行動とみなす（同，第 3 章）．オンライン推定による手法では，状態行動の局所予測モデルの更新および，ゴール到達可能性の二つの基準で初期は 1 状態ではじまり，随時状態を分割更新していく手法も提案され，実機で検証されている（同，第 4 章）．オンラインなので環境変化（例えばボールサイズが 2 倍に変化）にも対応可能である．

これらの手法はモデルフリー型から始まって，局所予測モデルなどを構築することで，モデルベースの学習法に変遷してきたともみなせる．この先駆者は Sutton による Dyna-Q [16] で，通常の Q-学習から得られるサンプルを用いて，世界モデルを構築し，これを用いて Q-学習して方策を精緻にし，それに従った方策による結果をまた世界モデルの更新に利用するやり方で，世界モデル構築と方策を分けた考え方である．この延長が深層学習をベースとした状態表現学習（SRL: State Representation Learning）とみなせる [14]．SRL の目的は，真の状態に基づく直接教示を避け，エージェントの行動や観測空間におけるその帰結に関する情報を用いることで，単純な特徴抽出ではなく，身体を有して環境と相互作用するエージェントによる状態表現を獲得することである．これは，認知発達ロボティクス [17] の核である「身体性」の概念が示す点であり，Lesort et al. [14] は能動学習や人工好奇心と関連付けている．

図 3 (a) に SRL の一般モデルを示す．円で示したのは観測可能で四角で示しているのが隠れた状態である．a_t, o_t, r_t は時刻 t における行動，観測信号，報酬を表し，$\tilde{s}_t$ は，真の状態でアクセス不能である．解くべき課題は，$\tilde{s}_t$ が未知の下で，以下のマッピングを学習することとされている [5]．

$$s_t = \Phi(o_{1\ t}, a_{1\ t-1}, r_{1\ t})$$

一般にニューラルネットワークを用いて，図 3 に示すように，オートエンコーダの一種である VAE (Variational Auto Encoder) などを利用した状態から観測信号の復元 (b)，順モデルの学習 (c)，逆モデルの学習 (d)，事前知識つきのモデル (e) などを組み合わせ，それぞれについて損失関数を規定し，最小化として学習が行われる．一例として，Finn, et al. [15] による視覚運動学習で使われた深層空間オートエンコーダを図 4 を示す．240*240*3 のカラー生画像から最終的に 32 次元の状態ベクトルに低減化されている．最後のたたき込み層は，空間ソフトマックを経由し，各チャンネルで活性度最大の点の位置を抽出する．最終的な復元画像はこれらの特徴点を使って再構成されている．彼女らは，これらの特徴点を用いて，局所線形モデルに基づく効率的強化学習により運動スキルを獲得し，PR2 を使ってプッシュ，ピックアップ，ハンギングなどの作業を実現した．

深層空間オートエンコーダに代表される状態表現学習では，順モデルも含めて，局所予測モデルが構築され，これを様々な局面で適用することで，より正確な世界モデルが出来上がる．これにより異なるタスク間でも利用可能であり，それらに共通する運動系列がプリミティブとして再利用可能である．局所予測モデルは，Uchibe et al. [12] もマ

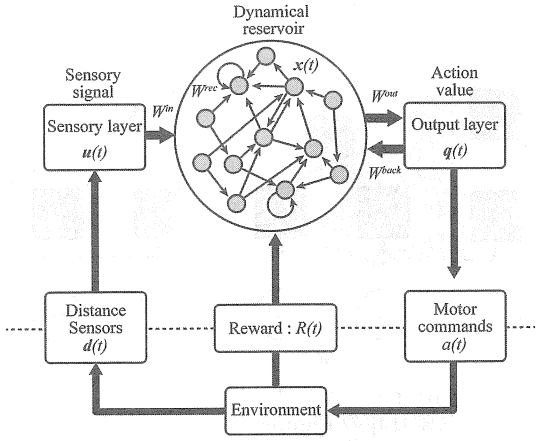

図 5 レザバーを用いた Q-学習（文献 [18] の Figure 3）

ルチエージェント強化学習で用いており，パッサーとシューターに交互に学習させて，シミュレーションでそこそこうまくなったら，実機での学習に切り替える手法をとっており，シミュレーションで学習過程はメンタルリハーサルとよばれている [4].

深層学習によるアプローチに対して，近年，レザバーを用いた計算手法が着目されている（例えば，文献 [19] など参照）．このレザバーと強化学習の組み合わせが Inada et al. [18] によって提案されている．**図 5** に，そのアーキテクチャを示す．センサ入力とモータ出力の間にレザバーを挿入し，報酬でネットワークにバイアスをかけることで，最初はランダムであった行動が徐々に，目標に到達する行動を学習していく．レザバーの役割は非明示的であるが，学習に必要な状態表現を獲得しているとみなせる．さらに，センサ信号の予測にも同様の形式でレザバーが利用可能で，環境モデル獲得過程とみなせる [20].

4.2 価値関数近似

価値関数近似手法は，強化学習において種々の領域への応用のスケールアップを可能にする重要なコンポーネントである．ニューラルネットワークを用いた関数近似が主流である，実環境応用では，Riedmiller et al. [21] が多層パーセプトロンを用いて価値関数近似し，バッチ型の教師あり学習の枠組みで様々なサッカーロボットスキルを獲得している．それらは，守備学習，インターセプション，位置制御，キック，モータの速度制御，ドリブル，ペナルティシュートなどである．キーアイデアはダイナミックプログラミング手法に従って，訓練用データセットを生成している点である．**図 6** (a) にバッチ型強化学習フレームワークのグラフィカルスケッチを，(b) に試合中の様子を示す.

4.3 報酬関数の推定

3.4 節で言及したように，吸収ゴールの設定で極値へのトラップは避けられるが，逆にすべての状態を経験しなければならず，永遠の経験は，実環境のロボットでは，疲弊や故障なども加わり現実的ではない．2 でカテゴリー化した教師あり学習は，Behavior Cloning とよばれているが，直接教示から間接教示，すなわち，模倣学習も教師あり学習の一つと考えられる．強化学習の観点からは，教示された状態行動系列を直接学ぶのではなく，(1) その裏にある報酬関数を推定し，(2) それに基づいて強化学習を進めることができる．前者の枠組は逆強化学習とよばれ，教示行動のみではなく，以外の状態行動にも対応できる.

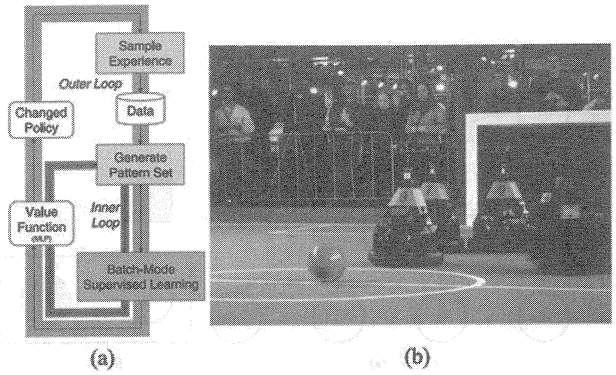

図 6 バッチ型強化学習フレームワーク（文献 [21] の Fig. 2 と 1）

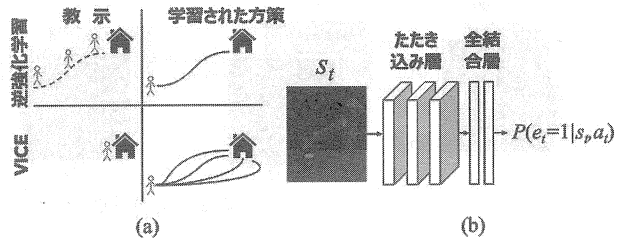

図 7 (a) 通常の逆強化学習と VICE の違い，(b) VICE の確率的グラフィックモデルパラメータ学習の枠組み（文献 [22] の Figures 1 と 2 を改編）

逆強化学習の枠組みでは，最大エントロピー原理に基づく手法や，さらに効率的にサンプリングや計算する手法が提案されており，解説 [5] で紹介されている．特に，敵対的生成モデル（GAN）の定式化を用いて，教示者と学習者の状態行動対を分類する識別器とそれを騙そうとする生成器の駆け引きとして強化学習と逆強化学習を組み合わせる手法は，深層学習が強化学習に深く入り込んで，深層学習が様々な学習に単なるツールとしてだけではなく，2 のカテゴリー化を超えて，新たな枠組みを提案しているように捉えられる.

ここでは，もう一歩進んで，逆強化学習で必要とされる教示者の状態行動列を必要とせず，ゴール状態のみを提示することで，報酬を推定する手法を紹介する [22]．これは，次節で述べるフルの実環境強化学習において，ゴール提示が非常に容易になることも強みである．Fu et al. が提案しているのは，VICE（Variational inverse control with events: a general framework for data-driven reward definition）とよばれる手法で，**図 7** (a) に示すように，通常の逆強化学習では，教示者の状態行動のみに依存した方策が獲得されるのに対し，VICE では，ゴールに到達可能な複数の経路が発見される．これは，ゴールのみを提示し，ゴールに至る事象の発生確率をデータから学習することで，可能にしており，この確率分布の関数近似に深層ニューラルネットワークが利用されている．同図 (b) にその枠組みを示す.

学習者のゴールは，積算報酬を最大化することではなく，将来どこかの点で一つかそれ以上のイベントが起こる確率を最大化することである．これは，制御をグラフィカルモデルの推定問題とし，報酬を事象生起変数とみなすことで，逆強化学習を一般化している．象徴的には，ゴール状態の画像を見せるのみで学習が効率よく進む．

5. 実世界強化学習の例

4.3節で紹介したVICEを含めて，UC Berkeleyのグループは Googleとも連携して，多くの実ロボットの深層強化学習の業績をあげている．そのなかで，実環境でフルにロボット学習に取り組んだ例を紹介する．Zhu et al. [6]は，実環境で，人間が介在しないで，最初からフルにロボットがタスクを学習するための三つのポイントを指摘し，それらについての解決策を提案している．それらは，

(1) 人が介在しないリセット機構：シミュレーションでは，一エピソード終了後，初期位置に容易に戻せるが，実ロボット学習では，実験者がいちいち手動で戻す場合が多く，労力がかかっていた．加えて，初期位置の分散が小さい傾向があり，これにより探索範囲が狭まっていた．これを解消するために，彼らは，ランダム化された摂動制御器（randomized perturbation controllers）を提案し，初期位置の分散を大きくとることで，結果として探索範囲を広めている．

(2) オンボードのセンサ情報のみに依存した観測：グリッドワールドは完全観測可能な環境設定だが，実世界は，ロボットのオンボードのセンサ情報のみが利用可能で，時間的にも空間的にも部分観測問題を含む．特に画像情報は自己位置視点に依存した高次元生データであり，VICE [22]などの手法を利用することで，生画像を直接，ゴール状態として与えることができる．また，状態表現学習も同様に，高次元データの低次元化に一役買っている．

(3) 報酬工学（設計者が意図的に報酬を設計）に依存しない報酬関数：報酬関数の設計は，これまで設計者が事前知識を生かした手作りや，入力が画像データの場合，完全なコンピュータビジョンアルゴリズムを用いた物体認識をベースにするなど，設計者の介入が重く，ロボットが自律的に報酬関数を推定する手立てが望まれていた．上と同様にVICEやVAE（Variational Auto Encoder）を用いて，高次元生画像を対象にして，ゴールをイメージで提示し，成否識別器を学習させることで，人間の介入なしに，ロボット自ら報酬関数と等価な情報を獲得し，タスクを成功させる方策を獲得する．

上記を組み込み，人間の介入を最小限に留めたシステムの概要を**図8**に示す．彼らは，実ロボット強化学習（Real Robot Reinforcement Learning）の略としてR3Lと称している．**図9**に，学習後の実際の実験の様子と結果を示す．上段がビーズ操作で，下段がバルブ旋回である．異なる初期状態から始まり，ゴールに到達している様子が窺える．今後の課題として，提示サンプルの複雑度，より複雑なタスクでの最適化と探索の困難，センサと機器作動のノイズなどがあるとしている．

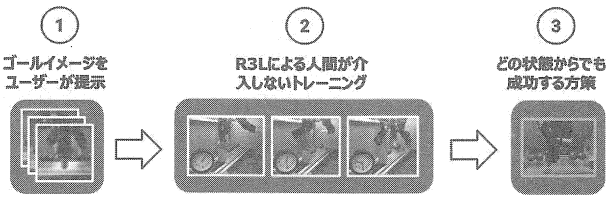

図8 R3Lの枠組み：(1) ユーザがゴールイメージや途中のサンプルイメージを提示し，(2) 人間の介入なしにロボットがトレーニングを続け，(3) 学習最後には，どの初期位置からでもゴールに到達する方策を獲得する．（文献[6]のFigure 1を改編）

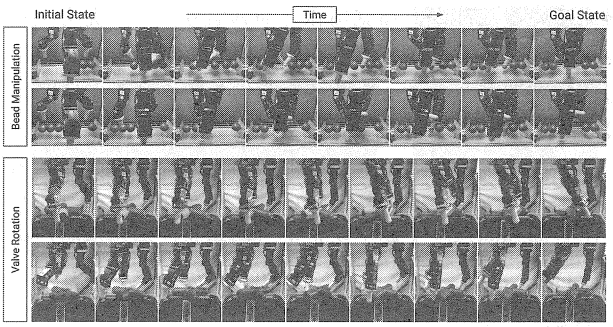

図9 R3Lによる実ロボット強化学習：ビーズマニピュレーション（上）とバルブ旋回（下）．（文献[6]のFigure 9）

強化学習と直接かかわらないが，汎用ツールとなりつつある深層学習技術を用いた実ロボット学習法として，深層予測学習法（Deep Predictive Learning）が提案・実験されている[23]．深層強化学習との対比では，状態表現学習に対応するのが，予測で生成できる画像のエラーから自己組織化される注意機構を用いて，状態空間を圧縮していることである．エンコーダ，再帰部，デコーダからなるアーキテクチャの再帰部において，運動と注意点を統合しており，運動に関係する画像部分のみ予測し，リアルタイムの行動生成を可能にしている．

6. おわりに

ロボティクスにおける強化学習の意味，位置づけ，応用の際の課題，解決法について最近のトピックを拾ってみた．実環境応用への課題そのものは，以前とあまり変わらないが，アプローチには，深層学習とその周辺技術の応用の効果が多大であると同時に，このことにより，これまでの機械学習のカテゴリーである，教師の有無や模倣学習の捉え方を組み直した感があり，今後もこの傾向は続くと想定される．これらの詳細に関しては，本特集号（日本ロボット学会誌 第39巻，7号 強化学習特集号）の記事を参照されたい．

Zhu et al. [6]の論文査読過程でも示され，彼らも認識しているのは，より複雑なタスク，例えば歩行を始めとするよりダイナミックな動きに関しては，アクチュエータなどの身体構造を考慮する必要があり，安全性の課題も含めて，よりメタな学習構造が必要と察せられる．

4.1節の後半で示したレザバーによる強化学習[18]では，移動ロボットのシミュレーションであったが，物理レザバーの観点[19]からは，ソフトロボティクスとの結合も有望視されており，推進しているNEDOプロジェクト†の成果も期待される．

ロボットの自律性の観点からは，そもそも報酬系を考えるうえでは，動機づけや好奇心の課題があり[24]，認知発達ロボティクス[25]の観点からは，より根源的な自己の概念や主体感との関連も追求したい課題である[26]．例えば，痛覚に関しては，筆者は情動や共感の発達に重要な要素と考えている[27]が，痛覚情報を制御信号として捉えた強化学習の枠組みも提案されており[28]，このような課題に深層強化学習がどのような解法を提供してくれるかに関しては，今後の動向を見守りたい．

謝　辞　本稿は，JST，CREST（JPMJCR17A4）津田プロジェクト，RISTEX HITE 稲谷プロジェクト，および国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託業務（JPNP16007）の結果得られたものである．

参　考　文　献

[1] 牧野貴樹，澁谷長史，白川真一（編）：これからの強化学習．森北出版，2016.
[2] 浅田稔（編著）：RoboCupSoccer: ロボットの行動学習・発達・進化．共立出版，2002.
[3] 浅田稔："実環境におけるロボットの学習・進化的手法の適用と課題"，計測と制御，vol.38, no.10, pp.650–653, 1999.
[4] J. Kober, J. Andrew Bagnell and J. Peters: "Reinforcement learning in robotics: A survey," The International Journal of Robotics Research, vol.32, no.11, pp.1238–1274, 2013.
[5] 長井隆行，堀井隆斗：強化学習とロボティクス．電子情報通信学会誌，vol.103, no.12, pp.1239–1247, Dec 2020.
[6] H. Zhu, J. Yu, A. Gupta, D. Shah, K. Hartikainen, A. Singh, V. Kumar and S. Levine: "The ingredients of real world robotic reinforcement learning," International Conference on Learning Representations, 2020.
[7] 浅田稔，國吉康夫：ロボットインテリジェンス．岩波書店，2006.
[8] R.S. Sutton, A.G. Barto and R.J. Williams: "Reinforcement learning is direct adaptive optimal control," 1991 American Control Conference, pp.2143–2146, 1991.
[9] T. Minato, Y. Yoshikawa, T. Noda, S. Ikemoto, H. Ishiguro and M. Asada: "Cb2: A child robot with biomimetic body for cognitive developmental robotics," Proc. of IEEE/RAS 7th Int. Conference on Humanoid Robots, pp.557–562, 2007.
[10] 浅田稔：浅田稔のAI研究道．近代科学社，2020.
[11] Y. Chebotar, A. Handa, V. Makoviychuk, M. Macklin, J. Issac, N. Ratliff and D. Fox: "Closing the sim-to-real loop: Adapting simulation randomization with real world experience," 2019 International Conference on Robotics and Automation (ICRA), pp.8973–8979, 2019.
[12] E. Uchibe, M. Asada and K. Hosoda: "Cooperative behavior acquisition in multi mobile robots environment by reinforcement learning based on state vector estimation," Proc. of IEEE Int. Conf. on Robotics and Automation, pp.1558–1563, 1998.
[13] M. Asada, S. Noda, S. Tawaratumida and K. Hosoda: "Purposive behavior acquisition for a real robot by vision-based reinforcement learning," Machine Learning, vol.23, pp.279–303, 1996.
[14] T. Lesort, N.D. Rodríguez, J.-F. Goudou and D. Filliat: "State representation learning for control: An overview," CoRR, vol.abs/1802.04181, 2018.
[15] C. Finn, X.Y. Tan, Y. Duan, T. Darrell, S. Levine and P. Abbeel: "Deep spatial autoencoders for visuomotor learning," Proc. of IEEE International Conference on Robotics and Automation, pp.512–519, 2016.
[16] R.S. Sutton: "Integrated architectures for learning,planning,and reacting based on approximating dynamic programming," Proc. of Workshop on Machine Learning-1990, pp.216–224, 1990.
[17] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino and C. Yoshida: "Cognitive developmental robotics: a survey," IEEE Transactions on Autonomous Mental Development, vol.1, no.1, pp.12–34, 2009.
[18] M. Inada, Y. Tanaka, H. Tamukoh, K. Tateno, T. Morie and Y. Katori: "A reservoir based q-learning model for autonomous mobile robots," The 2020 International Symposium on Nonlinear Theory and Its Applications (NOLTA2020), pp.213–216, 2020.
[19] 中嶋浩平，田中琢真，青柳富誌生："ダイナミクスによる情報処理――レザバー計算の最近の発展"，日本物理学会誌，vol.74, no.5, pp.306–313, 2019.
[20] M. Inada, Y. Tanaka, H. Tamukoh, K. Tateno, T. Morie and Y. Katori: "Prediction of sensory information and generation of motor commands for autonomous mobile robots using reservoir computing," The 2019 International Symposium on Nonlinear Theory and Its Applications (NOLTA2019), pp.333–336, 2019.
[21] M. Riedmiller, T. Gabel, R. Hafner and S. Lange: "Reinforcement learning for robot soccer," Autonomous Robots, pp.55–73, 2009.
[22] J. Fu, A. Singh, D. Ghosh, L. Yang and S. Levine: "Variational inverse control with events: a general framework for data-driven reward definition," Neural Information Processing Systems (NIPS), 2018, pp.8547–8556, 2018.
[23] H. Ichiwara, H. Ito, K. Yamamoto, H. Mori and T. Ogata: "Spatial attention point network for deep-learning-based robust autonomous robot motion generation," arXiv: 2103.01598[cs.RO], 2021.
[24] 浅田稔：内発的動機付けによるエージェントの学習と発達．牧野貴樹，澁谷長史，白川真一（編），これからの強化学習．森北出版，2016.
[25] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino and C. Yoshida: "Cognitive developmental robotics: a survey," IEEE Transactions on Autonomous Mental Development, vol.1, no.1, pp.12–34, 2009.
[26] 浅田稔："再考：人とロボットの自律性"，日本ロボット学会誌，vol.38, no.1, pp.7–12, 2020.
[27] M. Asada: "Artificial pain may induce empathy, morality, and ethics in the conscious mind of robots," Philosophies, vol.4, pp.38–47, 2019.
[28] B. Seymour: "Pain: A precision signal for reinforcement learning and control," Neuron, vol.101, no.6, pp.1029–1041, 2019.

†http://www.ams.eng.osaka-u.ac.jp/nedo-nmd/

浅田　稔（Minoru Asada）
大阪大学工学部教授，同大学大学院工学研究科知能・機能創成工学専攻教授，同大学先導的学際研究機構 共生知能システム研究センター特任教授（継続中）を経て現在に至る．知能ロボットの研究に従事（大阪大学名誉教授）．
（日本ロボット学会正会員）