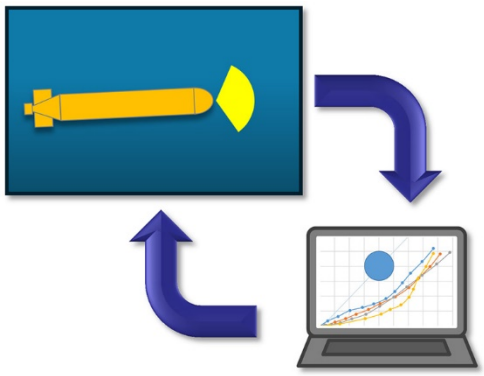
自律ビークル製品に向けた深層強化学習の応用技術

Advanced Technology of Deep Reinforcement Learning for Autonomous Vehicle



彌城 祐亮*1
Yusuke Yashiro

江口 和樹*2
Kazuki Eguchi

中川 陽介*1
Yosuke Nakagawa

深層強化学習は，反復試行を通じて自動的にデータを収集しながら学習し，最適な動作方策を求める手法として，近年注目を集め，適用例が増えている。目的地への移動や障害物の回避等，複数の目的を満たしながら自動制御で移動する自律ビークル製品へ適用するには，学習開始前に学習条件を適切に調整しなければならない課題がある。デジタルイノベーション本部では，ミッションである製品知能化の打ち手としてこの課題に取組み，深層強化学習を自律ビークルへより簡易に適用し，障害物回避動作の短縮等，効率的に製品性能を向上させるための改良手法を考案した。シミュレーション検証により，複数の障害物条件で効率的に障害物を回避する有効性を確認しており，その結果も併せて紹介する。

1. はじめに

自律ビークル製品は近年の省人化ニーズを満たすモビリティ製品として，陸・海・空を問わず社会に広がりを見せている(1)。自律ビークルがその役割を確実に果たすためには，複数の制御目標を満足するように，搭載される自動制御システム内の様々なパラメータを事前に調整する必要がある。近年は製品に要求されるタスクがより高度になり，また想定されない状況においても製品が安全な動作をする必要がある等，制御システムが事前に考慮すべきパラメータも増える傾向にあり，これらの調整がビークル運用で省人化を実現するためのハードルになっている。

一方で，情報技術の進展を背景に，深層学習をはじめとするデータ駆動技術が普及し始め，制御システムの自動調整にも適用が検討されている。特にデータの取得と最適な方策の学習を自動的に行う深層強化学習は，制御システムの構築・調整の一手法として期待されており，研究が進んでいる。しかし，この手法は内部の深層ニューラルネットワークに関するパラメータや，学習に最適なものを決める報酬関数等，調整項目が多い。そのため，単に適用しただけでは，制御システム内のパラメータ調整の手間が，学習に用いるパラメータの調整の手間に置き換わるだけになってしまうリスクがある。

本報ではこの問題を解決するために考案した，深層強化学習を適用しやすくする応用技術を紹介する。まず，調整が難しい報酬関数を，設計者の求める最適性を満たすように推定する必要があるが，これには，エキスパートデータを教師データとして，学習結果と同時に報酬関数も導出する手法である逆強化学習(2)を使用するとともに，複数の学習結果を選択しながら切替えて使用するセレクターを制御系に追加することで，深層強化学習の調整の手間を削減し，学習時に想定していない環境でも適切な学習結果を選択させ，効率的に制御を行う。本報では，自然科学研

*1 デジタルイノベーション本部 CIS 部　　*2 デジタルイノベーション本部 CIS 部　主席技師　工博

究・安全保障等幅広い分野で運用が広がっている水中自律ビークル（AUV：Autonomous Underwater Vehicle）の障害物回避を題材とし，シミュレーションで手法の有効性を検証している。本技術は当社グループの自律ビークル製品へ広く展開することで，情報化社会における高度で安全なモビリティの実現に貢献する。

2. 手法説明

図1に学習と，学習結果の制御系への適用の流れを示す。はじめに，逆強化学習を使用するための参照動作であるエキスパートデータを準備する。エキスパートデータは，ビークル製品に求める動作等から人為的に作成して良い。逆強化学習により，学習結果がエキスパートデータに近づくように，学習結果と併せて報酬関数も推定・自動的に導出されるため，報酬関数を調整する手間を省くことができる。例えば，ビークルが障害物を回避しながら目的地へ到達させる場合，目的地に到達するまでの時間を短くしようとすると，障害物の近くを通り，接触するリスクが増加するトレードオフがあるが，エキスパートデータと逆強化学習を用いることで，設計者が暗黙的に求めている時間と回避のバランスを推定させ，学習に反映できる。

また，深層強化学習は高い報酬値を得るように学習を進めていくが，最終的に得られる学習結果（方策）は用いた乱数条件によって異なり，また学習時とは異なる製品状況でも適しているとは限らないことから，報酬値が最大ではない結果も含め，複数の学習結果を制御に併用することで制御の頑健性向上を図る。本研究では，学習結果の選択には，複数の学習結果を導出しつつ，使い分けまで学習させるアンサンブル学習等のデータ駆動手法ではなく，セレクターと呼ぶシンプルな手法を使用する。セレクターは，使用者が事前に決めたランク付け関数を制御周期ごとに実行し，各学習結果を数値的に比較する。その時点で最適とした学習結果を使用し，次の制御周期では選択し直すため，自機位置と障害物の位置関係等の状況によっては，不適切となる学習結果を使い続けることを防ぐ。

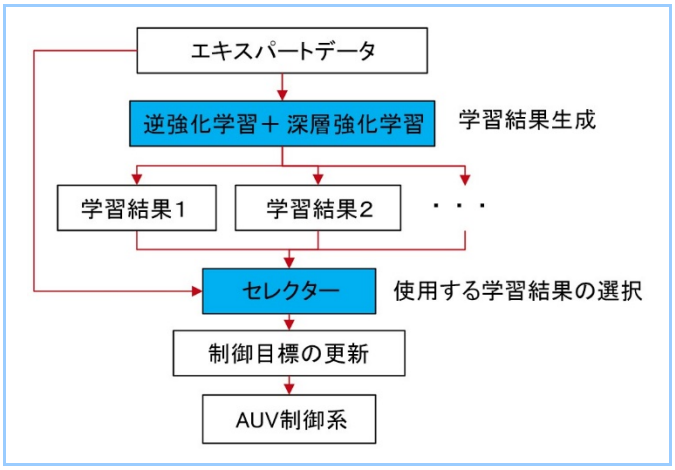


図1　提案手法の流れ

3. 障害物回避制御系

図2に複数学習結果とセレクターを用いた，AUV の障害物回避制御のブロック図を示す。各学習は正規化した空間で行い，その結果を位置データ点列として格納しておき，ビークルが障害物を検知した際の障害物までの距離に応じて，データ点列を都度スケーリングして使用する。AUV は，通常時は目標位置に向かって航行するが，障害物検知をトリガーにして，当初の目標位置から位置データ点列へ，一時的に目標位置を上書きする。学習結果を制御目標として使用することで，既存の制御系はそのまま使用できる。点列は1点目から使用され，制御周期ごとにセレクターのランク付け関数によって比較，切り替えられる。ランク付け関数の例として，本研究では各点列の点と当初の目標位置との距離を使用し，障害物回避をしながら目標位置に近づく点を選ぶよう

にした。選んだ点に到達すると，セレクターが次の点を選択し，これが繰り返される。点列の終端に到達した時点で回避が終了したとみなし，上書きされていた当初の目標位置へ戻す。

セレクターで同時に使用できるデータ数が多くなると，状況に応じて適切な選択肢を更に選びやすくなるが，データを用意するための学習に要する時間が増えるため，本報の例では4データとしている。この点については，逆強化学習・深層強化学習技術の今後の発展による学習時間の短縮が求められる。

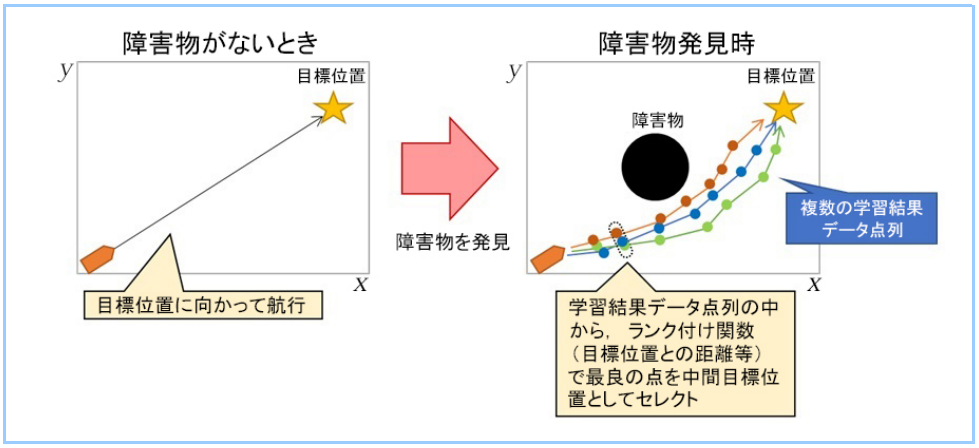


図2　提案手法による障害物回避制御

4. AUV シミュレーション検証

図3に AUV のシミュレーション設定を示す。本検証における AUV は船尾スラスタと舵を用いて旋回・航行し，慣性航法装置（INS：Inertial Navigation System）モデルにより自機位置取得を，前方ソーナーモデル1器のみを用いて障害物表面を検知し，AUV と障害物表面間の距離を取得させる。

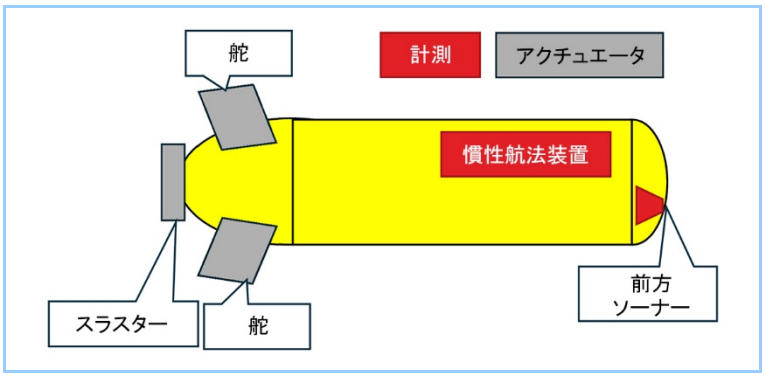


図3　シミュレータ検証用 AUV モデル

図4にシミュレーション結果を示す。提案手法の比較相手として，障害物を検知した場合，障害物の側方一定距離に一時的な目標位置を追加する手法を使用した。一方，提案手法では3つの学習結果回避軌道データと，逆強化学習に使用したエキスパートデータの計4データを使用した。

まず(a)学習時と同等の航路全長と障害物サイズを使用したベースケースを用いて，比較相手がやや早く目標位置へ到達するように，一時的な目標位置の追加距離を調整し，その後(b)評価ケースで2手法を比較している。評価ケースでは障害物を大きくしており，AUV にとっては学習時と異なる環境となる。

側方一定距離に一時的な目標位置を追加する場合，障害物の大きさが異なることで，ベースケースでは適していた回避距離では適切に回避できず，回避→回避終了→障害物再検知→回避と蛇行を繰り返し，704 秒かかっている。一方，提案手法では蛇行せず滑らかに回避経路を生

成・通過し，結果として500秒と目標位置により早く到達できている。これは航行中に，回避軌道として与えている位置データ点列の中から，都度目標位置へ最も近づくデータ点列をセレクターで選択して使用するため，不要な旋回が減少しているためである。(b)以外の条件を変えた評価ケースでも同様の優位性を確認しており，提案手法が異なる障害物に対して回避動作が大きく悪化することを防ぎ，頑健な回避手法となっているといえる。

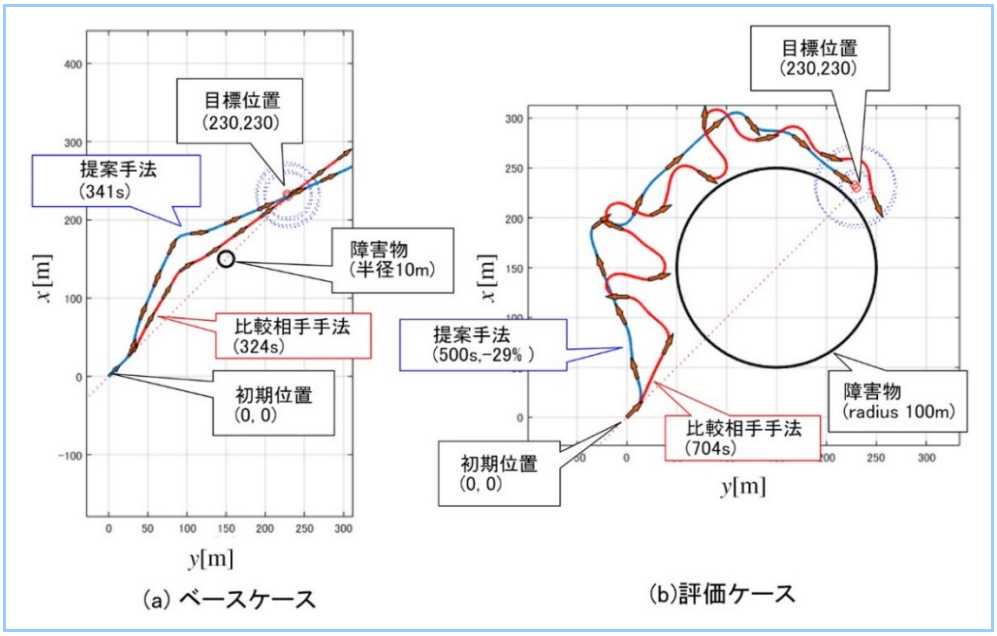


図4 検証シミュレーション結果

5. まとめ

複数の制御目標を持つ自律ビークル製品への適用を目指し，深層強化学習の調整の手間を省いて適用性を向上させる応用技術を考案し，シミュレーションで有効性を検証した。本手法によって調整が難しい報酬関数は逆強化学習によって自動推定され，求めるエキスパートデータに近い学習結果を得られると同時に，セレクターによって用いる学習結果を切り替えることで，学習時に考慮していない障害物等環境に対しても適切に制御される。これにより未知環境へも投入が求められる自律ビークル製品の，学習から制御までを含む運用を安全に進め，より高度なタスクの実現に貢献することが期待される。

残る課題としては，製品を運用開始した後に学習をやり直す必要がないように，どの程度の学習結果のバリエーションをそろえるか，製品の運用環境等と併せて事前に検討することが重要と考えられる。このバリエーションには学習時間が影響するが，強化学習アルゴリズムは日進月歩で発展を続けているため，より高速に学習できる手法がないか，学習アルゴリズムについても最新動向をフォローしていくことが求められる。今後もデジタルイノベーション本部体制を活かし，情報収集と製品技術開発を一元的に推進し，自律ビークル製品を使用するお客様の期待に応えていく。

なお，本手法は千葉大学と共同開発したものであり，同大学で協力いただいた荒井幸代教授，残間忠直准教授，研究室所属学生の方々に謝意を表します。

参考文献

(1) 江口 和樹ほか，AUV の低速･低高度制御，三菱重工技報 Vol.58 No.1 (2021)
(2) J. Fu, et al, Learning Robust Rewards with Adversarial Inverse Reinforcement Learning, ICLR 2018 (2018)
(3) Yashiro. Y, et al, Development of Applicable Reinforcement Learning Compensator Using Ranking Information for AUV, IEEE OCEANS 2022 (2022)