人追従走行ロボットにおける 深層強化学習を用いた適用モデルの簡易化

○池田 勇輝 (創価大学)　高橋 七海 (創価大学)　崔 龍雲 (創価大学)

1. はじめに

自動運転や物流支援など，ロボットを社会で導入することへの期待と共に，それに対応できる自律走行手法が期待されている．特に，人の支援を目的としたタスクの実現には，人を認識し追従走行する技術が求められている．例えば，ロボットが人を追従走行する手法として，複数のセンサを組み合わせた手法 [1] がある．これは，光を用いて各物体までの距離を測定する 2D-LiDAR [2, 3] と，パン・チルト回転機構を設けた RGB-D カメラを組み合わせることで，ロボットが人を認識し追従走行する精度を向上させた．このような人追従走行の手法により，宅配業務や，倉庫整理，買い物支援など，様々な分野での活躍が進んでいる [4].

これらの手法は，ロボットが人追従走行するために，人物追跡と走行制御を行っている．人物追跡では，ロボットに搭載されている各種センサを駆使することで，追従対象者の位置を検出し追跡する．また，走行制御では，検出した追従対象者の位置に合わせて経路を生成し，ロボットを経路通りに制御する．しかし，これらの処理は複数のセンサを組み合わせた上で，1 秒間に数十回の逐次処理を行っているため，走行制御までの計算に時間が掛り，人追従の性能が著しく低下する場合がある．

そこで，試行錯誤によって最適な行動を求める強化学習の導入により，これらの処理を簡易化することを考える．強化学習は，行動の主体となるエージェント (ロボット) が，環境との相互作用による試行錯誤を通じて，設定した報酬の総和が最大化するような行動を獲得する学習手法である．この強化学習を用いて，2D-LiDAR から得られるセンサ情報を入力とし，ロボットの行動を出力とするエージェント (方策モデル) を作成する．これにより，人追従走行手法の簡易化が期待できる．本研究では，2D-LiDAR から得られるセンサ情報を深層強化学習手法である SAC [5, 6] に適用することにより，ロボットによる人追従走行における簡易化手法を提案する．また，学習法の検証実験から本手法の有用性について報告する．

2. 深層強化学習による人追従走行の簡易化

本手法は簡易化手法の 1 つとして，2D-LiDAR から得られるセンサ情報と，ロボットや追従対象者の位置情報から，シミュレーション環境上で，深層強化学習である SAC による試行錯誤をすることで，学習効率を向上させる．本手法を実環境で動作させる場合は，シミュレーション環境で学習したエージェントを実環境に転移させることで実現可能である．これにより，実環境でもシミュレーション環境と同等の人追従走行が期待できる．

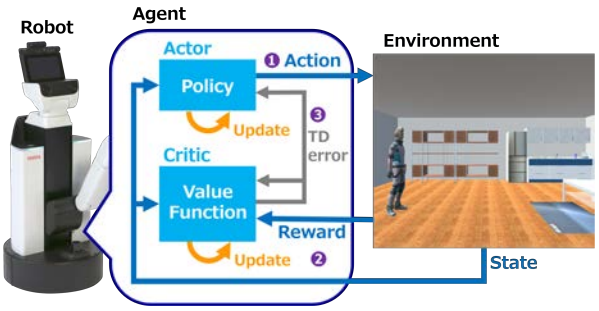
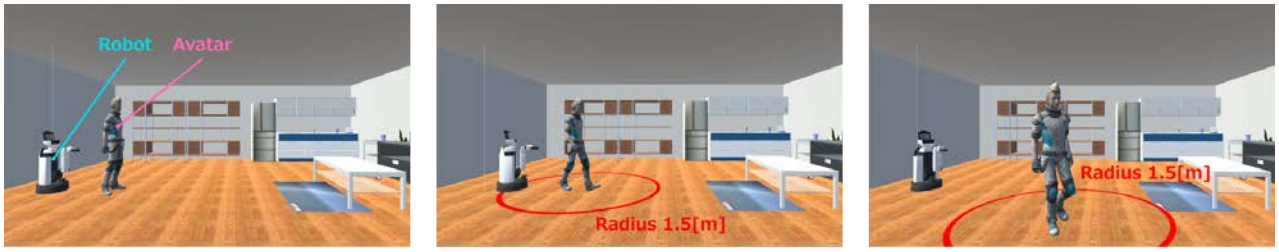


図 1　SAC を用いた人追従走行手法の構成

2.1 SAC を用いたロボットによる人追従走行手法

本手法では，追従対象者によってセンサ情報の差が少ない 2D-LiDAR を用いる．また，連続値の行動空間を利用する深層強化学習の手法である SAC(Soft Actor-Critic) をロボットの人追従走行に導入する．これは，関連研究 [7] において，RGB カメラから得られる画像が追従対象者の変化に対応できないことや，離散値の行動空間を利用した制御が追従走行に失敗する場合があることなど，様々な課題が見られたからである．本手法で用いる SAC は，Deep Neural Network で構成された，方策を表現するモデル (Actor) と価値を表現するモデル (Critic) を用いた手法であり，複雑な制御タスクに対して高いパフォーマンスを示している．連続値の行動空間を利用するため，細かい制御を必要とするロボットの分野で多く用いられている．本節では，2D-LiDAR と SAC を用いたロボットによる人追従走行の学習手順について，図 1 を用いて 3 つの段階に分けて説明する．

第 1 に，エージェント (Agent) は，現在の環境を観測し，それを踏まえた連続値の行動 (Action) を出力する．このときの本手法による処理のイメージを，図 1 の①に示す．本手法では，この環境を観測した状態 (State) として，ロボットに搭載されている 2D-LiDAR から得られるセンサ情報を用いる．エージェントは，Actor モデルに観測した状態であるセンサ情報を入力することで，ロボットの移動制御に用いる連続値の行動を出力する．本研究では，2D-LiDAR の距離情報に対して，カーネル内の最小値を取る Min Pooling を用いてデータサイズを圧縮したものを入力とする．この Min Pooling によって，ロボットとの距離が近い情報に注目できるため，安全性を保つと共に学習の効率を向上させることができる．また，移動制御に用いる連続値の行動には，並進運動 [m/s] と回転運動 [rad/s] の 2 つの運動を用いる．これにより，対向 2 輪型の移動ロボットと同等の移動制御を可能とする．そして，ロボットがエージェントによって出力された行動を実行することで，行動前に観測した状態に変化が生まれる．



(a) 初期状態　(b) 追従成功状態　(c) 追従失敗状態

図 2　SIGVerse を用いたシミュレーション環境

第 2 に，エージェントは実行した行動によって変化した状態から，行動結果の良さを表す報酬 (Reward) を受け取る．このときの本手法による処理のイメージを，図 1 の②に示す．強化学習は，設定した報酬の総和を最大化するように学習するため，人追従走行の精度向上には，この報酬の設計が重要となる．この報酬では，アバターとロボットとの距離が 1.5[m] 以下にある場合を追従に成功した状態とし，アバターとロボットとの距離が 1.5[m] より大きい場合を追従に失敗した状態とする．この条件を踏まえた報酬を式 (1) に示す．

$$reward = \begin{cases} 1.5 - distance & (distance \leq 1.5) \\ 0 & (1.5 < distance) \end{cases} \quad (1)$$

ここで，$reward$ は報酬，$distance$ はアバターとロボットとの距離 [m] とする．学習には，2.2 節で述べるシミュレーションの終了条件までの報酬の総和を用いる．

第 3 に，エージェントは報酬の総和を最大化するように，方策を表現する Actor モデルの最適化を行う．このときの本手法による処理イメージを，図 1 の③に示す．ここでは，価値を表現する Critic モデルから TD 誤差を算出し，Actor モデルと Critic モデルの最適化を行う．このように SAC は，方策を表現する Actor モデルと並行して，行動価値関数を近似する Critic モデルも同時に学習し，方策改善を行う．

以上で述べた 3 つの手順を，強化学習における 1 つのステップとする．また，このステップを 2.2 節で述べるシミュレーションの終了条件まで繰り返すことを，強化学習における 1 つのエピソードとする．本手法の学習は，このエピソードを一定回数繰り返すことで，ロボットによる人追従走行における最適な行動を学習する．

2.2　SIGVerse を用いたシミュレーション環境

強化学習は，一般的に数十万～数百万回の試行錯誤が必要となる．そのため，実環境より学習効率の高いシミュレーション環境を用いて，本手法の学習と検証を行う．本手法を実環境で動作させる場合は，シミュレーション環境で学習したエージェントを実環境に転移させることで運用する．このような運用事例は，参考文献 [8,9] で確認できる．

本研究で用いるシミュレーション環境として，国立情報学研究所にて開発されている SIGVerse [10] を用いる．SIGVerse は，Human-Robot Interaction(HRI) に関する必要な要素を，包括的に提供している研究プラットフォームである．このプラットフォームは，Robot Operating System(ROS) と Unity との間にリアルタイムブリッジング機構を開発することで，柔軟で再利用可能なシステムが実現できるようになっている．本研究では，SIGVerse による環境構築を行うことで，人追従走行における HRI を可能にする．

SIGVerse を用いて構築したシミュレーション環境を，図 2 に示す．本環境では，図 2 (a) に示す SIGVerse 上に用意されているアバター (Avatar) を追従対象者とする．このアバターは，シミュレーションが開始すると同時に，ランダムな方向に歩行する．また，そこのロボット (Robot) を用いて，アバターの追従走行を行う．このロボットは，エージェントから出力された行動を受け取り，動作する．エージェントは行動を学習するために，追従対象であるアバターとロボットの位置情報や，ロボットに搭載されている 2D-LiDAR の点群情報を受け取る．これらにより，ロボットによる行動とアバターを含む環境との相互作用が可能となる．

また，シミュレーション環境上でロボットによる試行錯誤を効率よく行うために，ロボットがアバターの追従に失敗した場合，シミュレーション環境を図 2 (a) に示す初期の状態に戻す処理を行う．このとき，2.1 節で述べた追従に成功した状態を図 2 (b)，追従に失敗した状態を図 2 (c) に示す．本研究では，追従に失敗した状態で，エージェントによる行動が 5 ステップ続いた場合，行っているシミュレーションを強制的に終了し，その環境を初期状態の図 2 (a) に戻す．また，学習を収束させるために，エージェントの行動回数であるステップ数が上限に達した場合も，同様にその環境を初期状態の図 2 (a) に戻す．

3.　学習の検証実験

本手法を用いた，ロボットによる人追従走行について学習の検証実験を行う．実験では，2.1 節で述べた SAC による人追従を，2.2 節で述べたシミュレーション環境に適用した簡易化手法を用いて学習する．そして，学習結果から，SIGVerse を用いたシミュレーション環境の評価と，本手法を用いたロボットによる人追従走行の学習精度について検証する．

3.1　検証実験の概要

実験では，本手法をシミュレーション環境上で 10 万ステップ繰り返し，ロボットによる人追従走行を学習する．ここでは，各エピソードでのステップ数の上限を 1000 ステップとして学習する．このとき，学習精度を明らかにするために，ロボットが 100 ステップ行動するごとに，その時点でのエージェントを評価する．この評価では，ロボットによる人追従走行を 3 エピソード行い，そのときの報酬和の平均を評価値として記録する．これにより，本手法の学習精度を確認する．

このとき，シミュレーションで用いる 2D-LiDAR は，

ロボットの正面から±120度の範囲に対して，961本のレーザによる距離情報を取得することができる．本実験では，2.1節で述べたMin Poolingで961本のレーザを30本まで圧縮する．また，本手法による実験を簡単にするために，Min Poolingの処理に加えて，2D-LiDARの情報量を削減したものを，エージェントへの入力として用いる．このとき，2D-LiDARから得られるセンサ情報に対して，ある閾値より大きいものを人追従走行に関係のない情報とすることで，センサの情報量を削減する．本実験では，この閾値を3.0[m]とし，センサ情報が3.0[m]より大きい値を0.0[m]に置き換える．

3.2 検証実験の結果

本手法で学習したエージェントの学習曲線を図3に示す．このグラフは，横軸に学習のエピソード数，縦軸にその報酬を表す．また，グラフ上の実線は100ステップごとの評価値，破線は全て評価値に対して最小二乗法を適用した近似直線による学習傾向を表す．このとき，試行錯誤を行っていない学習初期のエージェントによる1000ステップ分の評価値平均は0.8281，10万ステップの試行錯誤を繰り返したエージェントによる1000ステップ分の評価値平均は4.0621となった．このことから，学習前よりも評価値が3.2340程向上したため，人追従走行の精度が向上したことが言える．

また，図3のグラフ上の近似直線(破線)の傾きは2.9803×10^{-6}，切片は3.2081となった．このことから，学習初期から評価値が向上していても，10万ステップ全体における学習の中で，評価値の大部分が向上していないことが伺える．これは，多くのステップにてエージェントの学習ができず，近似直線の切片である評価値3.2081以上の精度で，ロボットが人追従走行を行えていない傾向にあることを意味している．このような結果となった原因の1つとして，エージェントに与えられる情報が少ないことが考えられる．本研究でエージェントに与えられる情報は，2D-LiDARから得られるセンサ情報のみであり，この情報から移動し続ける追従対象者の位置とその移動量を推測し，追従するのが困難であったように見受けられる．

さらに，多くの評価値は近似直線の切片である3.2081付近に分布しているが，26700，46900，89100エピソードでの評価値は25以上を取っている．これは，その他のエピソードでの評価値と比べて突発的であり，学習によって向上したものでないことが言える．このような結果になった要因として，シミュレーション環境上でロボットとアバターが家具や壁などに衝突して，動作が一時的に停止し続けたことが考えられる．これは，ロボットが人や障害物に衝突することを考慮していなかったため，このような状態が発生したと考えられる．

本実験から，学習初期と後期での評価値平均を比較することにより，SIGVerseを用いたシミュレーション環境で，ロボットによる行動と環境が相互作用し，強化学習可能であることが確認できた．また，評価値が学習前より3.2340程，向上したことが確認できる．しかし，近似直線の傾きが限りなく0に近いことから，多くのステップにて学習ができていない．このエージェントが，人を追従できたのは数秒間であり，これ以上の精度を向上させるには様々な改善が必要である．

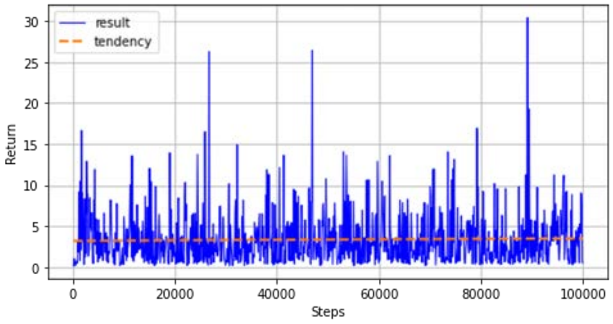


図3　学習曲線

4. まとめ

本研究では，2D-LiDARから得られるセンサ情報を強化学習に適用することで，ロボットによる人追従走行手法の簡易化を目的とした．そのために，ロボットに搭載されている2D-LiDARから深層強化学習によって最適化された行動をとり，人追従走行する手法を提案した．本手法による学習実験の結果，構築したSIGVerseによるシミュレーション環境で，人追従走行の強化学習が可能であることを示した．また，僅かではあるが，学習で評価値が3.2340程，向上したことを示した．しかし，多くのステップにて学習ができておらず，本手法の改善が必要である．これは，移動し続ける追従対象者を追従することに対して，エージェントに与えられる情報が少ないことが原因の1つであると考えられる．

このことから，本手法の学習精度の向上やその検討を行う必要がある．そのために，Actorモデルや Criticモデルの Network構成や，報酬の設計を再検討することが考えられる．また，ロボットが人や障害物に衝突することを考慮していないため，正確な人追従走行の精度検証ができていない．そのため，これらを考慮した報酬の再設計による本手法の確立が急がれる．

参　考　文　献

[1] 村上ら, LRFとパン・チルト回転機構を設けたRGB-Dセンサを用いた人追従走行ロボットの開発, ロボティクス・メカトロニクス 講演会, 2021.
[2] 北陽電機株式会社，北陽電機株式会社ウェブページ, http://www.ho- kuyo-aut.co.jp/, 参照日 2021-06-1.
[3] Velodyne Lidar, Velodyne Lidar Web Page, http://velodynelidar.com/, Visited on 2021-06-1.
[4] 株式会社ZMP，物流ロボの自動追従走行による公道での宅配利用提案について, https://www.zmp.co.jp/super-city/carriro-fd, 参照日 2021-06-01.
[5] T. Haarnoja, *et al.*, Soft Actor-Critic:Off-Policy Maximum Entropy Deep Reinforcement, *arXiv preprint arXiv:1801.01290*, 2018.
[6] T. Haarnoja, *et al.*, Soft Actor-Critic Algorithms and Applications, *arXiv preprint arXiv:1812.05905*, 2018.
[7] L. Pang, *et al.*, Efficient Hybrid-Supervised Deep Reinforcement Learning for Person Following Robot, *J Intell Robot Syst 97*, 2020.
[8] 有馬ら, 自律移動ロボットのための事前環境地図を必要としない深層強化学習を用いた動作計画, 第*33*回人工知能学会全国大会, 2019.
[9] K. Lobos-Tsunekawa, *et al.*, Point Cloud Based Reinforcement Learning for Sim-to-Real and Partial Observability in Visual Navigation, *arXiv preprint arXiv:2007.13715*, 2020.
[10] T. Inamura, *et al.*, *et al.*, SIGVerse: A Cloud-Based VR Platform for Research on Multimodal Human-Robot Interaction, *Frontiers in Robotics and AI*, 2021.