

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

Motion Planning with Deep Reinforcement Learning
without Using the Pre-Environment Map for Autonomous Mobile Robot

有馬 純平 黒田 洋司
Jumpei Arima Yoji Kuroda

明治大学
Meiji University

In this paper, we propose a learning-based motion planner for autonomous mobile robot which output continuous motion commands by referring from the scan data of 2D-LiDAR and the target point without using environmental map. We aim to use this method for a safe and efficient navigation system of an autonomous mobile robot with high generalization performance in urban environment. We show that a mapless motion planner can be trained through a deep reinforcement learning method in only simulator. In order to verify the effectiveness of the proposed method, collision avoidance and navigation performance are evaluated by directly applying the learned planner to the actual machine in the real world. As a result, we showed that we can obtain navigation performance equivalent to conventional method without using pre-environment map.

1. 緒言

近年、自動運転車を中心とした自律移動ロボットの研究開発が盛んに行われている。その中でも警備ロボットや配達ロボットなどの人間に近い環境で活動するロボットへの関心が高まっている。これらのロボットが自律移動を行うためには、環境認識、自己位置推定、経路計画の三要素が重要である。現在は、自律移動ロボットが事前に取得した環境のデータから作成した事前地図を所有し、その地図をもとに自律移動を行うことが一般的である。しかし、都市環境では事前に取得したデータと実際の環境が異なることが多い。このような場合、事前環境地図を利用したナビゲーションシステムでは、システムの安定性は実際の環境との差異に大きく左右されてしまう。そのため、ロバスト性の高い自律ナビゲーションを行うためには、事前環境地図に依存しないシステムが求められる。そこで、我々は、一般に公開されている電子地図から交差点を node、道を edge とした Edge-node map を作成し、これと通過すべき node のみを事前情報としたナビゲーションシステムの研究を進めてきた [澤橋 18]。しかし、このシステムで採用されている経路計画は、周辺の環境認識から、複数の経路を予測し、コストの少なくロボットのモデルを考慮した最適な経路を算出するため、環境の複雑さによって処理コストがかかる欠点がある。また、経路の予測時間、コストなどのパラメータの設計は環境に依存するものであり、我々の目指す環境に汎化性能のあるナビゲーションシステムには、従来のルールベースの経路計画は適していない。

本稿では、事前環境地図を必要せずに障害物を回避し目標位置に到達可能である動作計画を深層強化学習を用いて獲得する手法を提案する。事前環境地図なしの動作計画を実現するため、センサの入力と目標位置から動作指令を出力とする動作計画をシミュレーション上で学習させ、学習された動作計画は、直接実機に適用する。本稿では、学習された動作計画を、シミュレーション上と実世界において実験を行い本手法の有用性を示す。

連絡先: 有馬純平, 明治大学理工学部機械工学科, 〒 214-8571 神奈川県川崎市多摩区東三田 1-1-1, 044-934-7183,
arijun0307@gmail.com

2. 関連研究

近年、学習ベースで自律移動ロボットの行動を獲得する研究が多く行われている。

M. Pfeiffer らは、従来のルールベースによるナビゲーションを畳み込みネットワークで表現し、模倣学習させることによって、事前地図を用いずに従来手法と同等程度のナビゲーション性能を示した [Pfeiffer 17]。この研究では、2D-LiDAR をセンサ情報とし、畳み込みニューラルネットワークで動作計画をモデル化した。L. Tai らは、2D-LiDAR の 10 本のみを入力として、事前地図なしの動作計画を学習させた [Tai 17a]。この研究では深層強化学習のアルゴリズムである DDPG をベースとした手法を用いている。

上記の 2 つはセンサ情報として LiDAR を使用した研究であるが、カメラセンサを用いた研究も盛んである [Zhu 17][Tai 17b]。しかし、方策獲得に数百万回の試行錯誤が必要である強化学習は基本的にシミュレーションで行われるが、人工的なシミュレーション空間での画像と実世界での画像との違いによってモデルの転移に課題があり実用化は難しいのが現状である。一方学習させるネットワークの入力を低次元にすることによって、複雑な動的環境において障害物回避を行う研究も行われている [Fan 17]。この研究では事前地図を用いて建造物などの静的障害物を回避し、動的障害物は LiDAR とカメラを用いて位置と速度を推定している。これは事前環境地図の作成コストと認識の精度依存といった欠点が挙げられる。そこで本稿では、事前環境地図を必要とせず、パラメータ調整を行う必要のない End-to-end の動作計画を獲得することを目指す。

3. 提案手法

3.1 深層強化学習による動作計画

自律移動ロボットの動作計画は、センサから得られる情報と取るべき行動との関係を表すモデルを求め、新しい観測に対して直ちに動作指令を実行する必要がある。本稿では、式 1 に示すように、ある時刻 t におけるセンサ情報 x_t と目標位置 p_t から適切な動作指令 v_t を直接導出する方法を提案する。

$$v_t = f_\theta(x_t, p_t) \quad (1)$$

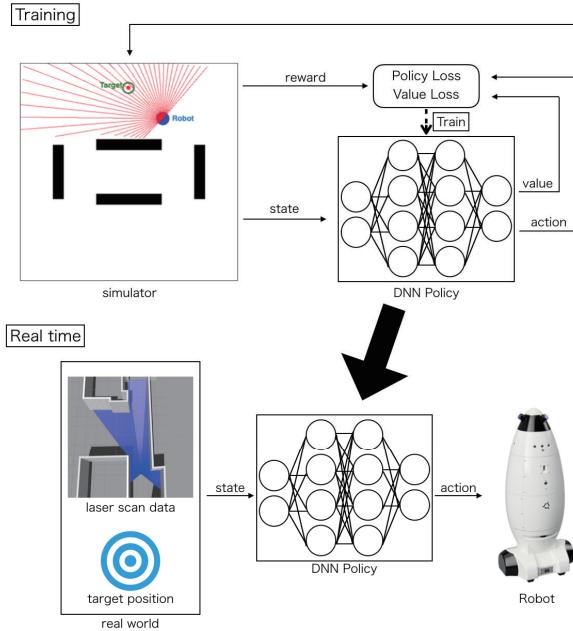


図 1: システム概要

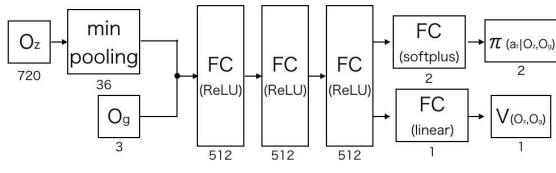


図 2: ネットワーク構造

ここで、ある時刻 t におけるロボットの状態から動作指令 v_t を出力する関数 f を θ でパラメタライズする。この θ はシミュレーション上で強化学習を行うことによって最適なものに調整する。本稿では、深層強化学習の手法として、Proximal Policy Optimization(PPO)[Schulman 17]を採用した。学習された方策は実空間へ直接適用する。本稿で提案するシステムの概要図を図 1 に示す。

3.2 ネットワーク構造

ロボットの状態と動作指令の間の End-to-end の関係を表すモデルは非常に複雑である。本稿では、この複雑モデルを万能関数近似器としてよく知られているディープニューラルネットワークで表現する。また、ロボットの状態として、2D-LiDAR から得られるセンサ情報とロボット座標系における相対的な目標位置を用いる。全体のネットワーク構造を図 2 に示す。

2D-LiDAR の情報は、ロボット正面から ± 90 度の範囲の 720 本のレーザの距離情報を表す。ここでは、720 本の距離情報を 36 本に圧縮したものを全結合層への入力とする。圧縮方法は、式 2 に示すように Min プーリングを用いる。

$$s_{l,i} = \min(s_{i \cdot k}, \dots, s_{(i+1) \cdot k}) \quad (2)$$

ここで、 k は 1D プーリングにおけるカーネルサイズを表し、本稿では $k=20$ とする。このように、Min プーリングしたものを受けとることによって、安全性を保つと共に学習の効率をあげることが出来る。また、シミュレーション上の限定的な環

境で学習を行うため、畳み込みニューラルネットワークによりセンサ情報から特徴量を抽出することはせず、Min プーリングを用いて実世界との相違を緩和する。ロボット座標系に対する相対的な目標位置は、距離と角度の正弦値と余弦値の 3 次元で表す。このセンサ情報と目標位置情報を合わせたものを 3 層の全結合層への入力とし、方策 π_θ と状態価値関数 V を出力する。本稿では、ロボットの動作指令値は連続値を扱うため、方策 π_θ は確率分布のベータ分布で表現する。方策へ入力する値は隠れ層の最後の出力に対して活性化関数の softplus 関数を用いた。

3.3 報酬設計

障害物に衝突せずに目標位置へ到達するというタスクをエージェントに学習させるために、本稿では式 3 に示す報酬関数を定義する。 λ^p , λ^ω は重み付け定数である。

$$R_t = r_t + \lambda^p r_t^p + \lambda^\omega r_t^\omega \quad (3)$$

ここで r_t は、衝突せず目標位置へ到達させるための報酬である。

$$r_t = \begin{cases} r_{\text{arrive}} & \text{if } d_t < c_d \\ r_{\text{collision}} & \text{if } \min_{s_l} < c_o \\ c_r(d_{t-1} - d_t) & \text{otherwise} \end{cases} \quad (4)$$

ここで r_t^p と r_t^ω は、目標位置との距離および方位を考慮した報酬であり、この項を入れることによりよりサンプル効率の良い探索を促す。

$$r_t^p = \begin{cases} r_{\text{position}} & \text{if } (d_{t-1} - d_t) = 0 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$$r_t^\omega = |\phi_t| \quad (6)$$

ここで d_t は時刻 t におけるロボットと目標位置との距離、 ϕ_t は時刻 t におけるロボットと目標位置との相対角度、 c_d と c_o はそれぞれ目標位置到達、衝突の判定距離である。

4. 実験

4.1 学習実験

提案する動作計画の訓練は、屋内環境を簡易的に模した OpenAI gym に準じた自作のシミュレーション環境で行う。目標位置は、各エピソードごとに障害物が存在しない場所にランダムに配置され、初期位置は環境の中心で初期方位はランダムとする。各エピソードの終了判定は、以下のいずれかとする。

- 目標位置に到達した場合
- 障害物に衝突した場合
- ステップ数が 500 ステップを超えた場合

また、学習の更新方法としては Adam を用いた。

提案手法による動作計画の訓練は、初期方策をランダムからはじめ、10 万エピソード行い学習が収束した。

4.2 評価

シミュレーション上で学習された動作計画の性能を評価するため、本稿では、学習環境とは異なる未知の環境においてシミュレーション上と実世界の双方で実験を行う。

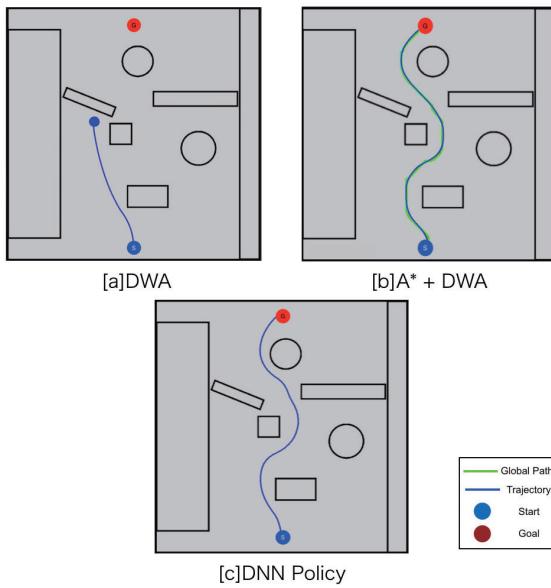


図 3: シミュレーション上での実験

表 1: シミュレーション上での実験の結果

| | A* + DWA | Proposed Method |
|-------------|----------|-----------------|
| Time[s] | 13.60 | 11.70 |
| Distance[m] | 10.51 | 10.43 |

4.2.1 シミュレーション上での実験

シミュレーション上での評価環境として、学習環境とは異なる障害物が複数置かれた $10m \times 10m$ の静的な仮想環境を作った。学習した動作計画の比較対象として、事前地図から A*アルゴリズム [Hart 68] を用いて大域経路を計画し、その経路に Dynamic Window Approach [Fox 97] で経路追従を行う従来手法と事前地図を用いず Dynamic Window Approach のみを用いた従来手法を採用する。提案手法では環境の地図は与えず目標位置のみを事前に与える。従来手法である 2 つと学習された動作計画の経路の結果を図 3 に示す。

図 3[a] より事前地図を与えず DWA のみで行う場合、行き止まりの道に入るとそこから抜け出すことができなくなってしまう。この理由としては、DWA は目標位置から離れるような行動を選択しないためだと考えられる。一方 A*を用いて計画された大域経路に DWA を用いて経路追従を行うとスムーズかつ障害物との適切な距離を保った動作を行えていることが図 3[b] からわかる。提案手法による軌跡は A*と DWA を組み合わせた手法と比較した場合僅かにふらつくことがあるが、似たような軌跡となり、障害物に衝突せず目標位置へ到達することができることがわかる。また、A*と DWA を組み合わせた従来手法と提案手法において上記の実験を 10 回行なったときの目標位置までの走行時間と走行距離の平均値を表 1 に示す。

結果から提案手法は事前地図を用いないが、シミュレーション上の静的な環境において従来手法と同程度のナビゲーション性能があると言える。表 1 では走行距離・走行時間双方で僅かに提案手法が優れているが、これは従来手法では大域経路を計画する際、コストの設計を安全に行なったおり、障害物からの距離を取るように経路を計画するため、距離が僅かに長くなつたと考えられる。

提案手法では、パラメータ調整をする必要がなくセンサ情

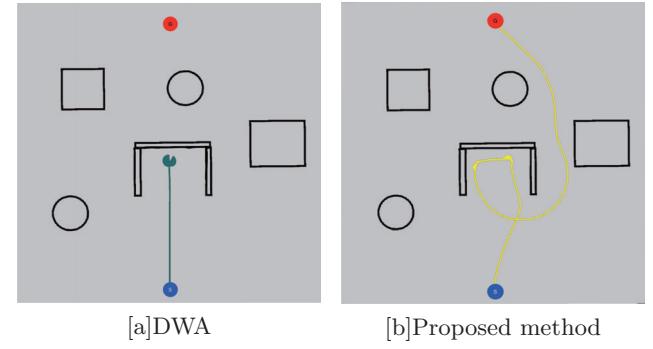


図 4: 袋小路環境における実験



図 5: 実環境における衝突回避実験

報から動作指令を出力することができる未知の環境において事前環境地図がなくとも柔軟に対応できるといった特徴があると言える。

また、上記の実験のように一般的に DWA による局所経路計画は袋小路などの局所解に陥りやすいという欠点がある。そこで初期位置と目標位置との間に袋小路が存在する環境で DWA による動作と提案手法による動作を比較した。図 4 にその結果を示す。

結果より、DWA による動作計画では、袋小路があるような環境で、そこから抜け出すことはできないが、提案手法では、抜け出し最終的に目標位置に到達することができる。提案手法では学習時にその場に留まることに負の報酬を与えていたため、目標位置が与えられている限り動き続けることを学習したからだと考えられる。

4.2.2 実世界での実験

シミュレーション上で学習された動作計画を、実機に搭載し実世界での評価を行なった。評価は、提案手法による障害物回避実験と、ナビゲーション実験の 2 つを行なった。

障害物回避実験では、実機として Roomba をプラットフォームとし、センサには 2D-LiDAR を用いた。

本実験では、ロボット座標系における目標位置の算出には、AMCL[Thrun 05] による自己位置推定を用いた。図 5 に実際に行なった走行テストの様子を示す。

このような目標位置までに複数障害物が存在する環境において学習済みモデルを直接実機に適用させた場合の軌跡を図 6 に示す。

結果から、シミュレーション環境ではない未知の環境において、シミュレーションのみで学習したモデルを実世界においても直接適用し衝突回避を行うことができる。これがわかる。

次に、ナビゲーションの実験について述べる。この実験では、実機として SEQSENSE 社から提供された SQ-2 を用いた。ナビゲーション実験は明治大学生田キャンパス第二校舎 D 館 1F において、図 7 に示す 3 点を通るように 1 周する実験を行なった。

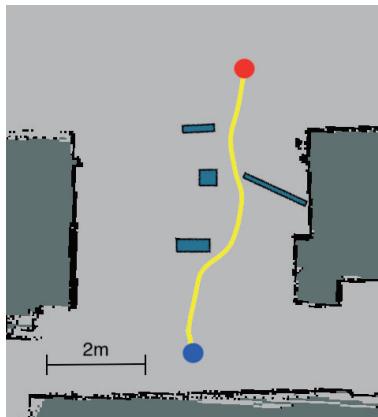


図 6: 衝突回避実験の軌跡

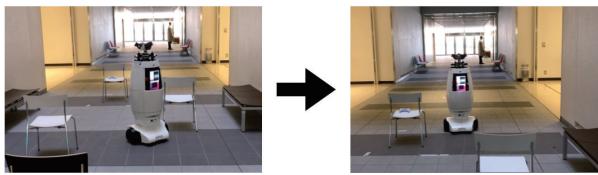


図 7: 実環境におけるナビゲーション実験

本実験で用いた SQ-2 のセンサは 3 次元点群を出力するので、それを 2 次元に圧縮し、学習済みのモデルへの入力とした。自己位置推定には衝突回避実験と同様に AMCL を用いた。実験の結果ロボットの結果を図 8 に示す。

結果より学習環境とは異なる環境においてナビゲーションを行うことができた。学習環境は $10m \times 10m$ の環境であったが、目標位置を数 $10m$ 離れたところに設置してもそこへ到達可能であることがわかった。また、目標位置へ最短距離の壁沿いに進むわけではなく、道の中央を走行することが確認できる。これは、報酬設計に明示的に組み込んだわけではないが、衝突回避を学習する上で、できるだけ障害物から離れるように学習したからだと考えられる。

5. 結言

本稿では、深層強化学習のアルゴリズム PPO を用いて、自律移動ロボットのためのセンサの入力と目標位置から連続値の動作指令を出力とする動作計画の構築手法を提案した。提案した手法により学習された動作計画は、シミュレーション上の未知の環境において、事前地図を用いずに従来手法と同程度のナビゲーション性能を示した。また実世界では、シミュレーション上のみで学習したモデルを直接適用することで、静的環境において障害物回避とナビゲーションを実現した。以上より提案手法の有用性が示せたと言える。

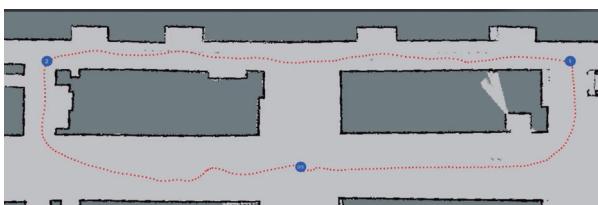


図 8: ナビゲーション実験の軌跡

謝辞

本研究の一部は、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、次世代人工知能・ロボット中核技術開発事業による支援を受けた。ここに篤く御礼申し上げる。

参考文献

- [澤橋 18] 澤橋遼太, 細田佑樹, 町中希彰, 山崎亮太, 定國裕大, 草刈亮輔, 黒田洋司, 「Edge-Node-Graph 及び分岐点検出に基づく道なり走行ナビゲーションシステムの開発」, 第 23 回ロボティクスシンポジア, 静岡県, 2018, pp.67-72
- [Pfeiffer 17] Pfeiffer, M., Schaeuble, M., Nieto, J., Siegwart, R., and Cadena, C.: From perception to decision:A data-driven approach to end-to-end motion planning for autonomous ground robots, in International Conference on Robotics and Automation (ICRA), pp. 1527-1533 IEEE (2017)
- [Tai 17a] Tai, L., Paolo, G., and Liu, M.: Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation, in International Conference on Intelligent Robots and Systems (IROS), pp. 31-36 IEEE (2017)
- [Zhu 17] Zhu, Y., Mottaghi, R., Kolve, E., Lim, J., Gupta, A., Fei-Fei, L., and Farhadi, A.: Target-driven visual navigation in indoor scenes using deep reinforcement learning, in International Conference on Robotics and Automation (ICRA), pp.3357-3364 IEEE (2017)
- [Tai 17b] Tai, L., Zhang, J., Liu, M., and Burgard, W.: Socially Compliant Navigation through Raw Depth Inputs with Generative Adversarial Imitation Learning, arXiv preprint arXiv:1710.02543, 2017.
- [Fan 17] Chen, Y. F., Everett, M., Liu, M., and How, J. P.: Socially aware motion planning with deep reinforcement learning, arXiv preprint arXiv:1703.08862, 2017.
- [Schulman 17] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O.: Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347. 2017.
- [Hart 68] Hart, P., Nilsson, N., and Raphael, B.: Formal Basis for the Heuristic Determination of Minimal Cost Paths, IEEE Transactions on Systems Science and Cybernetics, 1968.
- [Fox 97] Fox, D., Burgard, W., and Thrun, S.: The dynamic window approach to collision avoidance, Robotics Automation Magazine, IEEE, vol. 4, no. 1, pp. 23-33, 1997.
- [Thrun 05] Thrun, S., Burgard, W., and Fox, D.: Probabilistic Robotics pp. 250-261, MIT Press, 2005.