Vol.2022-DPS-190 No.27
Vol.2022-CSEC-96 No.27
2022/3/11

複数ポリシー切替による複雑な環境に適用可能な自律移動ロボットナビゲーション手法

天野 加奈子[1]　加藤 由花[1]

概要：近年，多様な動的環境に適用することを目的に，深層強化学習を用いた自律移動ロボットのナビゲーション手法が提案されている．しかし，学習される行動はシミュレーション環境に依存するため，現実世界に直接適用することは安全面および効率面で最適ではない場合がある．本稿では，人と空間を共有する自律移動ロボットが，人と障害物が混在する複雑な環境下において安全かつ効率的に目的地まで移動することを目的に，深層強化学習（DRL）手法を含む複数の行動決定方法を切り替えるナビゲーション手法を提案する．ここでは，新たにリセットポリシーを導入し，ロボット周辺の非占有領域の面積を用いて危険性の高い状況を判別することで，従来手法の課題である狭い環境での振動や衝突の回避を目指す．今回，深層強化学習手法単体を用いる場合と 3 種類の行動決定方法を切り替える場合においてナビゲーション実験を行い，環境中に静止障害物が存在するとき，切り替え手法の方が成功率が高いことを明らかにする．また，DRL・効率・安全の 3 つのポリシーの切り替え手法は，効率面で DRL ポリシー単体に劣ることや元々狭い環境では十分な成功率を得られないことを示す．

1. はじめに

近年，労働力不足の解消や感染症防止の観点から非対面・非接触のサービス提供が求められていることを背景に，サービスロボットへの期待が高まっている．サービスロボットは，空港や博物館，商業施設といった様々な人間が存在する環境において，清掃や運搬等のサービスを提供する．このような人間の生活圏内で活動する移動ロボットは，未知の障害物や歩行者に対応し，多様な環境下で安全かつ効率的に移動することが求められる．

これまで，動的な環境を対象とした自律移動ロボットのナビゲーションについて，様々な研究が行われている．アプローチの一つは，歩行者の経路を明示的に予測し，その予測結果を用いてロボットの行動計画を行うというものである．しかし，人が多く密集した環境では，いずれの行動も危険性が高いと判断しロボットが身動きを取れなくなる「Freezing Robot Problem」（FRP）が発生することが知られている．FRP の解決にはエージェント同士の協調行動を考慮することが有効であると考えられている [1]．また，歩行者の経路予測では歩行者の検知・追跡を行う必要があるが，ロボットに取り付けたセンサのみを用いる場合，混雑した環境ではデータの欠損等によって十分な精度が得られない可能性がある．これに対し，混雑した環境や実環境における様々な不確実性に対応することを目的に，深層強化学習（DRL）を用いてエージェント同士の協調行動を明示的もしくは暗黙的にモデル化するナビゲーション手法 [2] [3] が研究されている．一般的に，強化学習では，シミュレータ上に構築された学習環境の中でエージェントが試行錯誤を繰り返し，行動を学習する．そのため，意図した通りの行動が学習されているという保証がなく，現実世界に直接適用した場合，安全面および効率面で最適ではない行動を選択する可能性がある．これに対し，Sathyamoorthy ら [4] は FRP に対処するためのアルゴリズム Frozone を考案し，人の密度によって Frozone と DRL ポリシーを切り替えるアプローチを提案している．このアプローチは，密度が一定以下のとき Frozone を用いることで行動に保証性を持たせているが，Frozone は保守的な行動を生成するため，不要な回避が行われる場合がある．また，密度が高い場合は DRL ポリシーをそのまま利用するため，シミュレータと実世界のギャップに対処できない可能性がある．このほかに，複数の行動決定手法を切り替えるアプローチとして，Fan ら [5] は障害物との最短距離によって DRL ポリシー，DRL ポリシーの入出力を制限する Safe DRL ポリシー，PID 制御を切り替える手法を提案している．この手法は移動障害物がある程度多い環境であっても機能するが，密度が高い環境や狭い環境ではロボットの振動が起こ

[1] 東京女子大学 大学院理学研究科
Suginami, Tokyo 167–8585, Japan

Vol.2022-DPS-190 No.27
Vol.2022-CSEC-96 No.27
2022/3/11

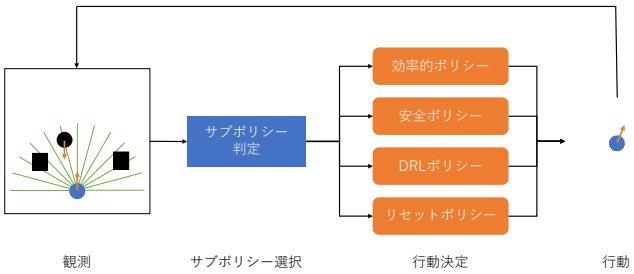

図 1 提案手法の概要

りやすい．

本稿では，密度が高い環境や狭い環境を含む様々な実環境において，安全かつ効率的なナビゲーションを行うことを目的とし，新たな切り替え判定方法とリセットポリシーを導入した複数ポリシー切り替えアプローチによるナビゲーション手法を提案する．具体的には，切り替えの判定にロボット周辺の移動可能な領域の面積を用いることで，ロボットが動けなくなる可能性が高い状況を判別し，DRL ポリシー，効率ポリシー，安全ポリシー，リセットポリシーという 4 つのポリシーを切り替えて様々な状況に適したナビゲーションを行う．リセットポリシーは，ロボット周辺の非占有領域が広くなる場所を探し，潜在的な危険性がある状況を脱するためのものである．これにより，既存手法では振動や速度の低下が起こりやすい密度が高い環境や狭い環境においても安定したナビゲーションを行うことことを目指す．提案手法の概要を**図 1** に示す．今回，複数ポリシーを切り替えるアプローチの有用性を検証するため，シミュレーション上の移動障害物と静止障害物が混在する環境においてナビゲーション実験を行う．具体的には，DRL ポリシー単体を用いる場合と DRL・効率・安全の 3 種類のポリシーを切り替える場合のナビゲーション性能を比較する．

2. 関連研究

本稿では，DRL ポリシーを含む複数ポリシーを切り替えることで多様な環境に適用可能なナビゲーション手法を提案する．本章では，関連研究として深層強化学習を用いたナビゲーション手法と状況に応じて複数のポリシーを切り替えるナビゲーション手法について述べる．

2.1 深層強化学習によるナビゲーション

近年，エージェント間の相互作用や実世界の不確実性を考慮した柔軟な行動計画の実現を目指し，深層強化学習を用いたナビゲーション手法が多く研究されている．深層強化学習ナビゲーションはエージェントレベルとセンサレベルのアプローチに分けられる．エージェントレベルのアプローチ [2] は，各歩行者の位置や速度を用いて明示的に各エージェント間の相互作用を学習に取り入れることができ，無駄の少ない効率的な移動が可能である．一方で，移動エージェントの検知・追跡モジュールを別途用意しなければならず，その精度がロボットの行動決定に影響を及ぼす．センサレベルのアプローチ [3] は，センサ計測値からエージェントの制御コマンドに直接マッピングする End-to-end のポリシーを学習するアプローチである．これは，実世界の不確実性にもロバストに対処することが可能である．エージェントレベルのアプローチに比べて最適な行動を学習することが難しいが，センサデータの表現の工夫 [6] や段階的な学習 [3] によって性能を向上させる取り組みが行われている．

また，一般的に，強化学習はモデルの学習に膨大な時間を要し，学習途中では危険性のある行動を選択する可能性がある．そのため，学習はシミュレータ上に構築された学習環境で実行される．このシミュレータ上の環境と実世界では，障害物の配置やエージェントのモデル等の差異が存在する．そのため，シミュレーションで学習したモデルを直接実機に適用した場合に適切に動作しないことがあり，この問題は Sim-to-Real ギャップと呼ばれる．この問題に対処するため，Rusu らによるプログレッシブネットアーキテクチャ [7] や Sadeghi らによる訓練セットのレンダリング設定を高度にランダム化する学習方法 [8] が提案されている．2.2 節で紹介する Fan らのハイブリッド制御アプローチ [5] は，複数のポリシーを切り替えることで上記の問題に対処しようとするものである．

本稿の提案手法においても，DRL ポリシーの利点を生かし，欠点を他のポリシーで補うため，複数ポリシーを切り替えるアプローチを採用する．今回，環境中のエージェントの検知が不要で混雑した環境にも対応でき，センサ構成の制限が少ないことから，DRL ポリシーとして Long らによる 2D-LiDAR を用いた End-to-end のアプローチ [3] を利用することを前提とする．

2.2 複数ポリシー切り替えアプローチ

複数の行動決定手法を状況に応じて切り替えることで，DRL のデメリットをカバーしようとするアプローチが研究されている．Sathyamoorthy ら [4] は FRP に対処するためのアルゴリズムである Frozone を考案し，人の密度によって Frozone と DRL ポリシーを切り替えるアプローチを提案している．このアプローチは，密度が一定以下のとき Frozone を用いることで，DRL 手法の行動に保障を持たせられないという欠点をカバーし，密度の高い時はロバストな行動を生成できるという DRL 手法の利点を活かしたものである．しかし，Frozone は回避する範囲を必要以上に広く取り，非効率な行動を選択する可能性がある．また，密度が高い状況では，DRL ポリシーを直接利用するため，シミュレータと実世界のギャップに対処できない．Fan ら [5] は，よりロバストなナビゲーションを実現する

ため，障害物との最短距離に応じて DRL ポリシー，DRL ポリシーの入出力を制限する Safe DRL ポリシー，PID 制御を切り替える手法を提案している．この手法は，移動障害物がある程度多い環境では単一の DRL ポリシーよりも高いナビゲーション性能を示し，最高速度や大きさが異なるエージェントが存在する環境にも対応することが可能である．しかし，密度が高い環境や狭い環境ではロボットの振動が起こりやすいことが課題として挙げられる．

本稿では混雑した環境等を含む多様な環境において安全性と効率性を高めるため，複数ポリシーを切り替えるアプローチのナビゲーション手法を提案する．シナリオ判定の際にロボット周辺の非占有領域の面積を用いることで，より効果的に危険性の高い状況を判別し，危険性が高い状況の回避や FRP からの脱出を試みるリセットポリシーの導入によって，従来手法の課題である狭い環境での振動や衝突を回避することを目指す．

3. 準備

本章では，提案手法を構成する要素技術を紹介する．3.1 節では深層強化学習について説明し，3.2 節では効率ポリシーとして用いる PID 制御について説明する．

3.1 深層強化学習

今回，提案手法を構成するサブポリシーの DRL 手法として，Long らが提案した End-to-end のマルチロボットシステムのための分散型衝突回避ポリシー [3] を用いる．本節ではこの手法における問題設定や学習アルゴリズムについて説明する．なお，我々のナビゲーション手法はマルチロボット環境に限らない一般の動的環境を対象とするが，この分散型衝突回避ポリシーは各ロボットは自分以外のロボットと位置等の情報を共有する必要がないため，ロボット以外の移動物体が存在する動的環境においても利用可能である．Long らはポリシーを学習するためのマルチシナリオ・マルチステージ学習の枠組みを提案しており，ポリシーはその枠組みの中で方策勾配に基づく強化学習アルゴリズムにより，シミュレータ上の環境で複数のロボットに対して同時に学習される．

まず，マルチロボット衝突回避問題を部分観測マルコフ決定過程（Partially Observable Markov Decision Process; POMDP）として定式化している．N 台のロボットのモデルは全て，半径 R を持つ円形の非ホロノミック差動駆動ロボットであることを前提としている．i 番目のロボット（$1 \leq i \leq N$）は，各時刻 t において，環境の状態 s_i^t から，確率的に観測 o_i^t を得る．そして，障害物に衝突せずに現在位置 p_i^t からゴール g_i へ向かうための制御指令値 a_i^t を決定し，実行する．強化学習では，ロボットは行動によって報酬 r_i^t を獲得し，それによって行動全体で得られる報酬の総和の期待値を最大化するように方策 π の更新を行う．

本手法では，各ロボットの観測 o^t は $o^t = [o_z^t, o_g^t, o_v^t]$ のように 3 つの要素で構成される．o_z^t はロボットに取り付けられた 2D レーザスキャンの測定値，o_g^t はロボットのローカル極座標系におけるゴールの座標，o_v^t はロボットの速度である．$o_z^t \in \mathbb{R}^{3\times512}$ は，角度範囲 180 deg，最長距離 4 m のセンサで観測される 512 本のレーザスキャンの 3 ステップ分で構成される．制御指令 $a^t = (v^t, w^t)$ は，ロボット前方の速度 $v^t \in (0.0, 1.0)$ とロボット中心の角速度 $w^t \in (-1.0, 1.0)$ を組み合わせたものであり，次の観測 o^{t+1} を受け取るまでの時間 Δt 内で実行される．報酬関数は，

$$r_i^t = ({}^g r)_i^t + ({}^c r)_i^t + ({}^w r)_i^t \tag{1}$$

のように設計されている．右辺の各項はそれぞれ，ゴールへ近づくことに対する報酬，衝突に関する報酬，スムーズな行動に対する報酬であり，

$$({}^g r)_i^t = \begin{cases} r_{arrival} & \text{if } \|p_i^t - g_i\| < 0.1 \\ \omega_g \left(\left\|p_i^{t-1} - g_i\right\| - \|p_i^t - g_i\| \right) & \text{otherwise} \end{cases} \tag{2}$$

$$({}^c r)_i^t = \begin{cases} r_{collision} & \text{if } \left\|p_i^t - p_j^t\right\| < 2R \\ & \text{or } \|p_i^t - B_k\| < R \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$({}^w r)_i^t = \omega_w \left| w_i^t \right| \quad \text{if } \left| w_i^t \right| > 0.7 \tag{4}$$

と定義される．B_k は静止障害物である．各パラメータは $r_{arrival} = 15$，$\omega_g = 2.5$，$r_{collision} = -15$，$\omega_w = -0.1$ に設定されている．

強化学習アルゴリズムには，Proximal Policy Optimization（PPO）[9] をマルチロボットシステムに拡張したものを用いる．また，学習は 2 段階に分かれており，第 1 段階ではロボットのみ存在する環境で学習を行い，第 2 段階ではロボットの他に障害物が存在するような複雑さを増した環境で学習が行われる．

3.2 PID 制御

PID 制御はフィードバック制御の代表的手法であり，出力結果と目標値の差とその微分および積分を利用して制御を行う手法である．操作量 $u(t)$ は次の式によって定まる．$e(t)$ は目標値 $r(t)$ と出力値 $y(t)$ の偏差である．

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau)\, d\tau + K_d \frac{de(t)}{d\tau} \tag{5}$$

$$e(t) = r(t) - y(s) \tag{6}$$

右辺の各項は，出力と目標値の偏差，偏差の積分，偏差の微分に関する項である．パラメータは比例ゲイン K_p，積分ゲイン K_i，微分ゲイン K_d であり，これらを制御対象に合わせてチューニングする必要がある．

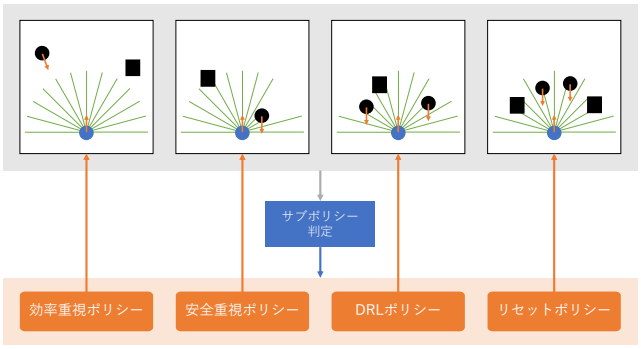

図 2 シナリオの分類

Algorithm 1 サブポリシー判定

```
if min(o_z^t) > R_safe or min(o_z^t) > ||p_i^t − g_i|| then
    Safe scenario (efficient policy)
else if min(o_z^t) ≤ R_risk then
    Risk scenario (safety policy)
else if S ≤ S_risk then
    Potential risk scenario (reset policy)
else
    Scenarios other than the above (DRL policy)
end if
```

4. 提案手法

本章では，我々が提案する，状況に応じて行動決定手法を切り替えるナビゲーション手法について述べる．4.1 節にて手法の全体像を説明し，4.3 節と 4.2 節ではシナリオ判定方法と選択可能なサブポリシーについてそれぞれ説明する．

4.1 概要

本手法は，図 1 に示すように，ロボットが観測した状況に応じてサブポリシーを選択し，その選択されたサブポリシーに従って行動するという一連の流れを繰り返すことでナビゲーションを行う．サブポリシーの選択フェーズでは，最適なポリシーが異なると考えられるシナリオを

- 障害物との衝突可能性が低い
- 障害物との衝突可能性が高い
- 潜在的な衝突可能性や FRP 発生可能性が高い
- 上記 3 つ以外

の 4 つに分類し，**図 2** のように，効率ポリシー，安全ポリシー，リセットポリシー，DRL ポリシーの中から各シナリオに適したポリシーを選択する．シナリオの判定には，ロボットと最も近い障害物の間の距離，およびロボット周辺の非占有領域の面積を指標として用いる．

4.2 サブポリシー

本手法において選択可能な次の 4 つのサブポリシーの機能について詳しく説明する．

- DRL ポリシー
- 安全ポリシー
- 効率ポリシー
- リセットポリシー

1 つ目の DRL ポリシーは，深層強化学習手法によって学習されるポリシーである．本手法では，Long らの 2D レーザスキャンデータを入力としてロボットの行動を直接出力する End-to-end のポリシーを前提とする．これは，移動物体の追跡モジュール等が必要なく，センサの観測誤差に強い．また，非協力的なエージェントも回避できるといった汎化性能が示されている．しかし，シミュレータ環境で学習された行動が実環境において最適であるという保証がないため，他のポリシーと組み合わせる必要がある．

2 つ目の安全ポリシーは，衝突の危険性が高い場合に，ロボットの速度を制限して衝突を回避するためのものである．Fan らのハイブリットアプローチの Safe DRL ポリシーにあたる．Fan らは DRL ポリシーの入出力に制限をかけたものを Safe DRL ポリシーとしており，本手法においても同様の方法を採用する．これは，未知の実環境に DRL ポリシーを適用するとき，安全性を確保する上で必要となる．ただし，このポリシーの頻度が多いと，ロボットの速度の低下によって効率的に移動できないことや，速度変化が多くなり周囲の人間にとって不自然な動きになることが考えられる．加えて，安全ポリシーに切り替わったとしても，立ち往生となって危険性の高い状況を脱することができない場合がある．そのため，安全ポリシーが必要とされる状況自体を回避することや立ち往生した場合にナビゲーションを継続するための行動が必要である．

3 つ目の効率ポリシーは，衝突の危険性が低い場合，高い速度を保ち直接ゴールへ向かうためのものである．Fan らのハイブリットアプローチの PID ポリシーがこれにあたる．本手法においても効率ポリシーとして，3.2 で紹介した PID 制御を用いるものとする．

4 つ目のリセットポリシーは，近いうちに安全ポリシーが必要となるような潜在的な危険性がある状況を回避するためのポリシーである．具体的には，(7) 式にあるロボット周辺の非占有領域 S が一定以上となる場所を探すよう行動する．これにより，安全ポリシーの説明で述べたような安全ポリシーが必要となる状況自体の回避や立ち往生からの復帰を実現する．

4.3 サブポリシー判定方法

図 2 にあるような 4 種類のシナリオの判定方法をアルゴリズム 1 に示す．潜在的な衝突可能性や FRP 発生可能性が高いシナリオの判定には，本稿で提案するロボット周辺の非占有領域の面積 S を用いる．ロボット周辺の非占有領域は，ロボットのレーザスキャンの範囲内で障害物が存在

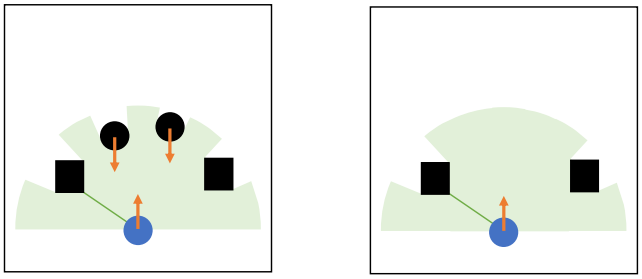

図 3 ロボットと障害物間の最小距離では区別できない状況の例．青い円はロボット，黒い円や四角は他の移動物体や静止障害物を表し，オレンジの矢印は移動方向を表す．緑線はロボットと最も近い障害物の間に引かれており，(a)，(b) で同じ長さである．薄い緑はロボットの観測範囲を表す．

しない領域のことを指し，**図 3** の薄い緑色の部分である．例えば，図 3 の (a)(b) は最も近い障害物との距離が同じ状況であるが，(a) はロボットが移動する方向に障害物が多く，(b) よりも衝突可能性が高い状況であると判断できる．このような状況を区別するため，ロボット周辺の非占有領域の面積 S をシナリオの判定の指標として用いる．ロボット周辺の非占有領域の面積 S は，レーザスキャンの本数を m，レーザスキャンの角度範囲を $[r_{min}, r_{max}]$ として，

$$S = \sum_{i=1}^{m} \frac{1}{2} o_{z_i}^t \sin\left(\frac{m}{r_{max} - r_{min}}\right) \qquad (7)$$

のように求める．ここで，$o_{z_i}^t$ は m 本のレーザスキャンのうち i 本目で計測された距離の値である．

5. 実験

本稿では，単一の DRL 手法を用いる場合と複数手法を切り替える場合のナビゲーション性能の違いを明らかにするため，シミュレータ上で実験を行った．今回，複数手法の切り替えに関しては，Fan ら [5] のように DRL ポリシー・安全ポリシー・効率ポリシーの 3 種類を切り替えるものを実装した．本章では，実験に関する各種設定および実験結果について述べる．

5.1 環境

本実験では，ロボット用ソフトウェアプラットフォームの ROS[*1] および 2 次元のロボットシミュレータの Stage[*2] を用いる．シミュレータ上に複数のロボットと静止障害物を配置して実験環境を構築し，ナビゲーション実験を行う．実験環境内の全てのロボットは，2D レーザスキャナセンサを搭載し，同一のナビゲーション手法によってそれぞれのスタート地点からゴール地点まで移動するものとする．センサの範囲は 180 deg，6.0 m とする．各ロボットは自分以外のロボットと位置等の情報を共有しないため，ロボットを他の障害物と明示的に区別することはない．そのため，ロボットをロボット以外の移動物体（人間等）と置き換えることができ，マルチロボット環境に限らず，より一般的な動的環境における衝突回避ナビゲーション問題として考えることが可能である．ロボットが衝突を起こした場合，ロボットはその場で停止し，障害物として残るものとする．

[*1] http://wiki.ros.org
[*2] http://wiki.ros.org/stage

5.2 指標

ナビゲーション性能を表す指標として，各手法において以下の項目を比較する．なお，今回は環境中にロボットが複数存在するが，各指標は環境中の全てのロボットのデータを用いて算出するものとする．制限時間は，実験環境の大きさや実際にナビゲーションを行なった様子から，100 s と設定した．

成功率 全試行中，1 度も衝突せずに時間内にゴールへ辿りついたロボットの割合．

衝突率 全試行中，他のロボットや静止障害物に衝突して停止したロボットの割合．

タイムアウト率 全試行中，衝突しないものの時間内にゴールできなかったロボットの割合．

平均移動時間 1 度も衝突せず時間内にゴールしたロボットがゴール到達までに要した時間の平均．

平均速度 1 度も衝突せず時間内にゴールしたロボットの速度の平均．

5.3 比較手法

今回，DRL ポリシー単体の場合と DRL ポリシー・安全ポリシー・効率ポリシーの 3 種類を切り替える場合のナビゲーション性能を比較する．DRL ポリシーは，3.1 節で説明した Long らの手法を用いる．

3 種類のポリシー切り替えは，アルゴリズム 1 から 5，6 行目の Potencial risk scenario の判定を除いたフローで行われ，これは Fan らのハイブリッドアプローチ [5] と同様の方法である．安全ポリシーは，Fan らのハイブリッドアプローチにおける Safe DRL ポリシーと同様に，DRL ポリシーの入力と出力を制限したものを用いる．入力に関しては，パラメータ p_{scale} によって，レーザスキャンに関する観測値 o_z^t を $\hat{o}_z^t = \frac{o_z^t}{p_{scale}}$ のようにスケーリングし，出力される速度は $[-v_{max}, v_{max}]$ の範囲にクリッピングする．また，今回の実装において，効率ポリシーではアルゴリズム 2 のように速度を算出する．ここで g_x，g_y は，ロボットのローカル座標系におけるゴールの x 座標と y 座標である．

各パラメータの値を**表 1** に示す．これらは，先行研究のパラメータの値を参考に決定した．

5.4 シナリオ

ナビゲーション実験を行うシナリオについて説明する．

Vol.2022-DPS-190 No.27
Vol.2022-CSEC-96 No.27
2022/3/11

Algorithm 2 効率ポリシーの実装

$v = v^t + 0.1$
if $v > v_{max}$ **then**
  $v \leftarrow v_{max}$
end if
$\theta = \arctan\left(\frac{g_x}{g_y}\right)$
if $\theta < -0.2$ **then**
  $w \leftarrow (v, -0.5)$
else if $\theta > 0.2$ **then**
  $w \leftarrow (v, 0.5)$
else
  $w \leftarrow (v, 0.0)$
end if

表 1 切り替え手法のパラメータ

v_{max}	0.5
p_{scale}	1.0
R_{safe}	1.0 [m]
R_{risk}	5.0 [m]

今回，静止障害物と移動障害物が混在する環境や狭い環境におけるナビゲーション性能を検証するため，次の 2 通りのシナリオを用意した．

円形 **図 4** のようにロボットを 10 台と静止障害物 9 個配置したシナリオ．各ロボットのゴールは，初期位置から円の中心を通った反対側である．

通路 **図 5** のような通路にロボットを配置したシナリオ．各ロボットのゴールは，初期位置から離れた通路の反対側であり，スタートからゴールまでの距離は全ロボットで同じである．

5.5 結果

それぞれの手法にて各シナリオで 10 回ずつ試行を行なった結果を**表 2** に示す．いずれのシナリオにおいても，切り替え手法の方が成功率が向上した．静止障害物が多い環境において DRL 手法の成功率が低下する理由としては，ロボットは障害物が移動することを過度に期待していると考えられる．切り替え手法を採用することで，障害物に衝突する前に速度を落とすことができ，回避するための時間に猶予が生まれる．そのため，DRL 手法に比べて成功率が向上していると考えられる．これは，切り替え手法の方が平均移動時間が長く，平均速度が遅くなっていることからも言える．密集した環境では安全ポリシーに切り替わる状況が多いため，慎重な行動によって成功率が上がるものの，ゴールまでに時間がかかる結果になっていると考えられる．

また，通路のシナリオにおける成功率は，切り替え手法の方が高いものの 75 % にとどまっている．これは，ロボットの視界が遮られることや元々移動できるスペースが少ないこと，ジグザグの形状の通路で他のロボットの行動が予想しづらいことが原因と考えられる．

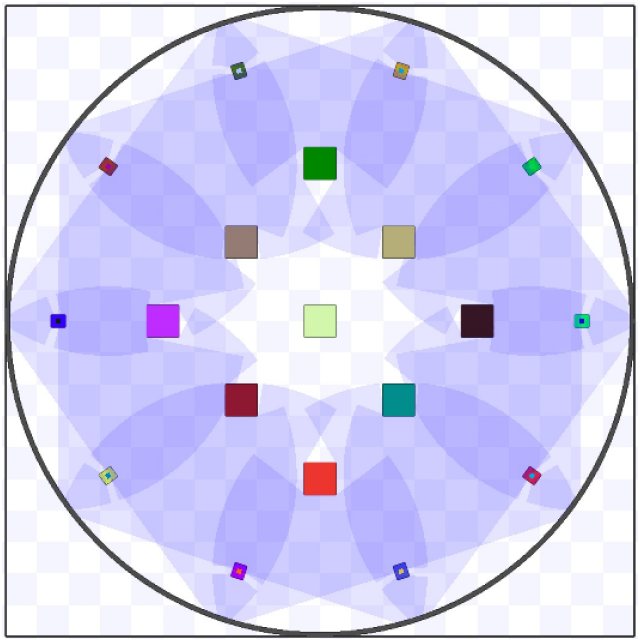

図 4 複数のロボットと障害物を配置した円形のシナリオ．中心に点（センサ）がある小さい四角がロボットであり，薄い青い部分はロボットのセンサ範囲を表している．濃いグレー部分およびロボット以外の図形は壁や障害物である．

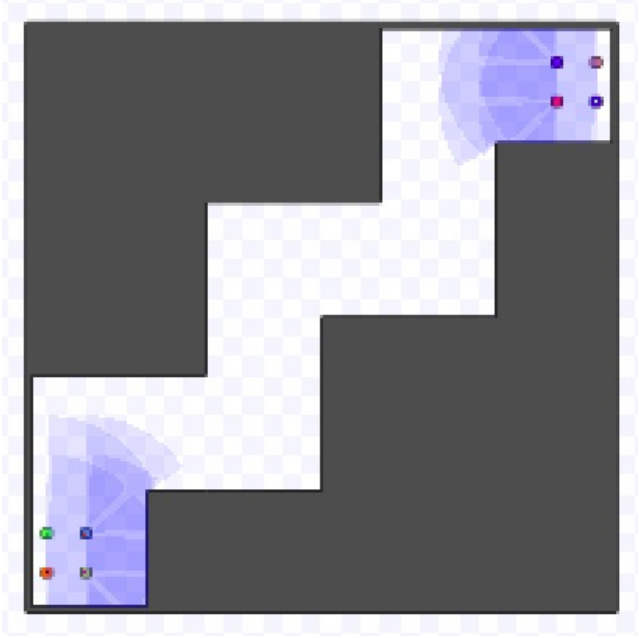

図 5 通路のシナリオ．中心に点（センサ）がある小さい四角がロボットであり，薄い青い部分はロボットのセンサ範囲を表している．濃いグレー部分は壁となっている．

今回の DRL・効率・安全ポリシーの切り替え手法では，円形のシナリオにおいて高い成功率を達成したが，より狭く不確実性の高い通路のシナリオでは十分な成功率が得られないことがわかった．また，円形のシナリオでは切り替え手法の成功率が高かった一方，移動時間や速度などの効

表 2 ナビゲーション性能の比較

指標	手法	シナリオ	
		円形	通路
成功率	DRL	0.69	0.58
	切り替え	0.96	0.75
衝突率	DRL	0.31	0.38
	切り替え	0.04	0.21
タイムアウト率	DRL	0.0	0.038
	切り替え	0.0	0.038
平均移動時間	DRL	21.09	40.51
	切り替え	26.15	49.30
平均速度	DRL	0.90	0.98
	切り替え	0.71	0.79

率面は従来の DRL 手法に劣っていた．これらは，危険性の高い状況の見落としや安全ポリシーの回数の多さが原因と考えられる．そのため，本稿で提案したシナリオの判定やリセットポリシーの導入がこれらの解決に有効であると考える．

6. おわりに

本稿では，複数のポリシーを切り替えることで様々な環境に適用可能なナビゲーション手法を提案した．実験では DRL ポリシー単体の場合と 3 つのサブポリシーを切り替える場合のナビゲーション性能を比較し，静止障害物が多い環境や狭い環境では，サブポリシーを切り替えることで成功率が向上することを示した．今後，提案手法におけるサブポリシー判定の実装やリセットポリシーのアルゴリズムの検討・実装を進め，提案手法の有用性の検証を行っていく．

参考文献

[1] Trautman, P. and Krause, A.: Unfreezing the robot: Navigation in dense, interacting crowds, *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, pp. 797–803 (2010).
[2] Chen, C., Liu, Y., Kreiss, S. and Alahi, A.: Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning, *2019 International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 6015–6022 (2019).
[3] Long, P., Fan, T., Liao, X., Liu, W., Zhang, H. and Pan, J.: Towards optimally decentralized multi-robot collision avoidance via deep reinforcement learning, *2018 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 6252–6259 (2018).
[4] Sathyamoorthy, A. J., Patel, U., Guan, T. and Manocha, D.: Frozone: Freezing-free, pedestrian-friendly navigation in human crowds, *IEEE Robotics and Automation Letters*, Vol. 5, No. 3, pp. 4352–4359 (2020).
[5] Fan, T., Long, P., Liu, W. and Pan, J.: Fully distributed multi-robot collision avoidance via deep reinforcement learning for safe and efficient navigation in complex scenarios, *arXiv preprint arXiv:1808.03841* (2018).
[6] Cui, Y., Zhang, H., Wang, Y. and Xiong, R.: Learning world transition model for socially aware robot navigation, *2021 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 9262–9268 (2021).
[7] Rusu, A. A., Večerík, M., Rothörl, T., Heess, N., Pascanu, R. and Hadsell, R.: Sim-to-Real Robot Learning from Pixels with Progressive Nets, *Proceedings of the 1st Annual Conference on Robot Learning* (Levine, S., Vanhoucke, V. and Goldberg, K., eds.), Proceedings of Machine Learning Research, Vol. 78, PMLR, pp. 262–270 (2017).
[8] Sadeghi, F. and Levine, S.: Cad2rl: Real single-image flight without a single real image, *arXiv preprint arXiv:1611.04201* (2016).
[9] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O.: Proximal policy optimization algorithms, *arXiv preprint arXiv:1707.06347* (2017).