Vol.2021-DPS-186 No.63
Vol.2021-CSEC-92 No.63
2021/3/16

歩行者の移動傾向を考慮した強化学習による自律移動ロボットナビゲーション

一色 春香[1]　天野 加奈子[1]　加藤 由花[1,a)]

概要：本稿では，人と空間を共有する自律移動ロボットが，安全かつ効率的に目的地まで移動するナビゲーション問題を対象に，強化学習を用いた経路計画手法を提案する．ここでは，ロボットと歩行者の相対的な位置関係，および対象となる歩行者の移動傾向（速度と移動方向）を学習時の状態に組み込むことで，歩行者の進路を妨げない経路の生成を実現する．シミュレータを用いた評価実験により，提案手法は，ポテンシャル法等の既存手法と比較し，ゴールまでの到達時間は長くなるものの，歩行者の進路を妨げず，衝突を回避する安全な行動を取ることを示す．

1. はじめに

近年，人と共存する自律移動ロボットに対する期待が高まっている．これらロボットの活用想定シーンは商業施設や空港，博物館と幅広く，提供するサービスも案内・清掃・運搬・警備など多岐に渡る．これらのシーンで利用されるサービスロボットは，人と空間を共有するため，安全性に配慮した動作が求められる．特に，歩行者が存在する動的環境下で行動する移動ロボットの場合，ロボット自身が歩行者を認識し，回避行動をとる必要がある．また，移動ロボットは環境内を自由に移動できるよう，自身に搭載したバッテリーを動力源とするのが一般的であり，不要な動作による電力消費を可能な限り抑えることが望ましい．そのため，安全かつ効率的な移動経路の計画は，移動ロボットにとって重要な研究課題になっている．

動的環境における移動ロボットの経路計画については，これまでも多くの研究開発が行われてきた．障害物とゴールにポテンシャル関数を定義し，その勾配により進行方向を決定するポテンシャル法 [1] や，ロボットを配備する環境下で計測した歩行者軌跡データを用いて人存在確率や人位置を予測することで，経路計画に利用する手法 [2], [3] などである．前者は，ロボットの近傍から得られる情報のみを利用するため，未知の環境においても障害物との衝突を回避する行動を策定できる一方，目的地から遠ざかる非効率な経路が生成される可能性がある．後者は，環境内の人の行動特性に従った経路を計画できる一方，ロボットを配備する環境ごとにあらかじめ歩行者軌跡データを収集する必要があり，導入できる環境が限られるという問題がある．

これらの問題を解決するために，本稿では，歩行者を回避しつつ効率的に目標位置に到達するロボットの行動を強化学習により獲得し，それを経路計画に利用する手法を提案する．強化学習は，行動の主体であるエージェントが環境の状態を観測し，何らかの行動により報酬を得るという一連の流れを繰り返すことで，得られる報酬の期待値を最大化する行動を獲得する手法である．本稿では，時間の経過と歩行経路への侵入に対して負の報酬を与えることで，安全性と効率性の双方を考慮した行動を学習する．環境に依存しない行動を獲得するために，状態として，通常利用されることの多い絶対座標系における歩行者・ゴールの位置座標の代わりに，ロボットに設定された相対座標系における位置座標を用いる．さらに，観測された歩行者の速度と移動方向を状態に組み入れることで，歩行者の移動傾向を考慮する．

本稿ではさらに，提案手法の有効性を検証するために，シミュレータを用いた評価実験を行う．ここでは，行動計画の安全性と効率を従来手法と比較する．また，選択する状態の影響を評価する．本稿の貢献は以下の 2 点である．

- 動的環境における強化学習を用いた経路計画を対象に，安全性と効率の双方を考慮した状態構成法を明らかにする．
- シミュレーション実験により，提案手法の安全性と効率を既存手法と比較することにより検証するとともに，学習時の状態として用いる項目の影響度合いを明らかにする．

[1] 東京女子大学
Tokyo Woman's Christian University, Suginami, Tokyo 167–8585, Japan
[a)] yuka@lab.twcu.ac.jp

2. 関連研究

本章では，歩行者等の移動障害物が存在する動的環境を対象としたロボット経路計画手法として，局所的経路計画に関する研究，歩行者経路予測と組み合わせた手法に関する研究の 2 種類について説明する．また，強化学習を利用したロボットの行動計画に関する研究についても紹介する．

2.1 局所的な経路探索手法に関する研究

移動障害物が存在する環境では，ロボットの近傍の情報を利用して，局所的な経路決定，移動を目標点到達まで繰り返す局所的経路探索手法がよく利用される．例えば，安全性を高めるため，移動障害物の進行方向に仮想的な障害物を想定することで，障害物の速度を考慮したポテンシャル法 [4] などが提案されている．

これらの手法は，突発的な移動障害物の出現に対応可能である一方，移動障害物の経路によってはロボットが不要な回避や待機を行う可能性がある．また，短い周期で経路を更新するため移動効率が悪化したり，急に向きを変えるなど，人間にとって親和的でない行動を取る場合もある．

2.2 歩行者経路予測と組み合わせた手法に関する研究

環境内の人の移動傾向を考慮しつつ，効率的に目標地点に到達するために，人移動軌跡データを取り入れた経路計画手法が提案されている．そのうちの 1 つに，人移動軌跡データから環境内の人の存在確率を導き，RRT*（Rapidly exploring Random Tree star）により交通量の多いエリアを避ける経路計画手法 [2] がある．この手法では，動的障害物と遭遇する可能性や経路を再計画するリスクを軽減することが可能である．しかし動的障害物の存在確率を表す Abundance map 作成のために環境内にセンサやカメラを設置する必要がある．また，環境地図や人の移動傾向が変化した場合はこの map を作り直す必要があり，利用できる場所が限られる．その他，動的環境において大域的経路探索を可能にするために，時空間グラフに障害物移動経路を不可侵領域として組み込む手法 [5] も提案されている．この手法は，施設案内のように人の移動経路が定まっている環境を前提としているため，人が自由に行き来するような環境では，歩行者経路予測など，他の手法と組み合わせる必要がある．

そのため，歩行経路予測と組み合わせた経路計画手法が提案されている．これらの手法では，環境内の人の移動傾向に合った歩行者経路をリアルタイムで予測し，時系列的な情報を経路計画に反映さることが可能である．特定の環境内の人移動データから歩行者移動モデルを生成することで歩行者経路を予測し，XYT 空間において経路計画を行う手法 [3] では，歩行経路予測に等速直線運動モデルを利用する場合に比べ，タスクの成功率（ゴールへの到達率）が向上することを示している．一方，ロボットの速度や回転などの制約が考慮されておらず，実ロボットへの適用には問題が残っている．LSTM（Long-Short Term Memory）を用いた人移動経路予測を取り入れた RRT*による経路計画手法 [6] では，従来の RRT ベースのアルゴリズムと比較し，探索時間や再計画回数を減らせることを示している．空間的な情報に加えて，時間的情報を反映できるため，ある瞬間の障害物の状態のみを考慮する場合と比べて，効率的な経路計画が可能である．

2.3 強化学習を用いた行動計画に関する研究

機械学習や深層学習の進展により，強化学習を取り入れた自律移動ロボットに関する研究も活発に行われている．強化学習は，ロボット自身の試行錯誤により目標達成のモデルを獲得するアルゴリズムであり [7]，未知の環境や変化する環境に適応した行動を獲得することが可能である．事前に大量の学習が必要であるが，学習終了後には高速に軌道生成が可能である．

ロボットの経路計画に関しては，学習結果を用いて経路探索アルゴリズムや深層強化学習の目的関数，報酬関数のパラメータを決定する手法が多く提案されている [8], [9], [10]．ロボットへの動作指令を強化学習により直接獲得する研究もある．静的環境においては，環境地図を事前に取得せずに動作計画を実現するため，深層強化学習を用いてセンサ情報と目標位置から動作命令を直接獲得する手法が提案されている [11]．

動的環境においては，移動障害物の次時刻遷移先の予測と実際の観測点の差異を評価して行動選択を行う予測型強化学習による行動計画が提案されている [12]．この手法は，特定の環境に依存しない回避行動獲得が可能であるが，衝突予測に基づいた危険度によってのみ行動を決定するため，危険度が同じ場合の行動優先順位を手動で設定しないと，目的地点に到達できないという問題がある．

本稿では，強化学習を用いることで，動的環境において適切な経路計画を実現することを目指す．特に，状態の定義方法について考察する．

3. マルコフ決定過程と Q 学習

まず，提案手法で用いるマルコフ決定過程と Q 学習 [13] について説明する．

3.1 マルコフ決定過程と動的計画法

マルコフ決定過程は，環境の取り得る状態の有限集合 $\mathcal{S}$ とエージェントがとる行動の有限集合 $\mathcal{A}$ によって定義される．時刻 t において環境中の状態 $s \in \mathcal{S}$ にあるロボットがエージェントに与えられた制御指令により行動 $a \in \mathcal{A}$ を実行し，時刻 $t+1$ において状態 $s' \in \mathcal{S}$ に遷移するときの状態遷移確率は $P(s' \mid s, a)$ と表される．このとき，エー

ジェントが環境から受け取る報酬 r も確率的に決定され，その期待値は $R(s,a,s')$ で表される．ここで，状態遷移確率，報酬の期待値はともにマルコフ性を持っており，時刻 t 以前の状態や行動履歴に依存しない．エージェントは何かしらの行動ルール（方策 Π）に従って行動 a を選択し，行動の評価値を最大化するような行動選択の規則を見出していく．評価値は遷移後の状態 s' における報酬と価値の和の期待値から成り，状態価値関数は，

$$V^{\Pi}(s) = E_{P(s'|s,a)}\left[R(s,a,s') + V^{\Pi}(s')\right] \qquad (1)$$

と表される．方策 Π は得られた状態価値関数を利用して改善することができる．ある状態 s において既存の行動 $a = \Pi(s)$ よりも高い評価値を得られる行動があれば，エージェントは行動を変えた方が良いことになる．このとき，状態価値関数の行動 a を変数として行動価値関数は，

$$Q^{\Pi}(s,a) = E_{P(s'|s,a)}\left[R(s,a,s') + V^{\Pi}(s')\right] \qquad (2)$$

と定義できる．行動の書き換えの手続きは方策改善と呼ばれ，行動価値関数 $Q^{\Pi}(s,a)$ を用いて，

$$\Pi(s) = \underset{a\in\mathcal{A}}{\operatorname{argmax}}\, Q^{\Pi}(s,a) \qquad (3)$$

と表す．全行動 $a \in \mathcal{A}$ の行動価値関数 $Q^{\Pi}(s,a)$ を求め，最大の行動価値を $V(s)$ に代入することを繰り返し，価値の更新が収束したときに得られる状態価値関数 V^* の示す方策が最適方策 Π^* となる．

3.2 Q 学習

Q 学習は，ロボットが行動をとった後に得られる情報から行動価値関数を更新していくアルゴリズムの一つである．すべての状態 $s \in \mathcal{S}$，行動 $a \in \mathcal{A}$ に対する行動価値関数の値 $Q(s,a)$（以後，Q 値と呼ぶ）を任意の初期値に設定して学習を開始し，エージェントの試行錯誤により，以下のように Q 値を更新していく．

$$Q(s,a) \longleftarrow (1-\alpha)Q(s,a) + \alpha\left[r + \max_{a'} Q(s',a')\right]. \quad (4)$$

ここで，α $(0 < \alpha < 1)$ は学習率と呼ばれ，α の値が大きいと遷移前の情報の減衰が早くなる．また，$\max_{a'} Q(s',a')$ は遷移先の状態 s' において一番価値の高い行動 a' を選んだときの Q 値を指し，このように行動を選択する方策はグリーディ方策と呼ばれる．グリーディ方策により学習するロボットは同じ状態で同じ行動を選択するに留まり，別の行動を試さないため，方策を改善することができない．そこで，基本的にはグリーディ方策だが確率 ε でランダムに行動を選択する ε-グリーディ方策がとられることが多い．

4. 提案手法

4.1 前提条件

本稿では，博物館，空港，大型商業施設の通路のように道幅が一定程度の広さを持ち，歩行者が通路内を自由に行き来する環境を対象に，自律移動ロボットが歩行者との衝突回避および目的地点到達を実現するための行動計画手法を提案する．ここでは，以下の前提条件を仮定する．

- 対向 2 輪型のロボット（駆動輪 2 つが車体を挟んで同軸についているロボット）を対象にする．これは，それぞれの車輪の回転速度を変えることで，その場での回転や曲線的な移動が可能なロボットである．
- ロボットへの制御指令は，前方方向への速度 ν [m/s] と中心の角速度 ω [rad/s] の組 $u = (\nu,\omega)^{\mathsf{T}}$ で与える．
- ロボットは環境地図を保有しており，自己位置推定が可能であるとする．
- 環境内の歩行者数は一人とする．
- 目標位置（ゴール）は既知とする．

4.2 手法の概要

提案手法の概要を **図 1** に示す．手法は Q 学習による学習フェーズと経路計画フェーズの 2 段階で構成される．Q 学習による学習フェーズでは，シミュレータ上の環境で人移動軌跡データを使用し，ロボットと歩行者，目標地点の位置関係や歩行者の速度，移動方向などの状態に応じてエージェントがロボットに与える制御指令を学習する．安全性と効率性に優れた行動を得るため，時間の経過と歩行経路への進入に対して負の報酬を設定し，目標位置に到達した場合と歩行者に衝突した場合を終端状態とする．経路計画フェーズでは，事前に学習フェーズで獲得した方策を参照することにより目標地点到達までの経路計画を行う．Q 学習および経路計画の手順を以下に示す．

- Q 学習
  (1) 歩行者の観測
  (2) 状態の更新
  (3) ε-グリーディ方策に基づく行動決定
  (4) 報酬の獲得
  (5) 状態 s と行動 a の組に対する Q 値の更新
- 経路計画
  (1) 歩行者の観測
  (2) 状態の更新
  (3) 行動決定
  (4) 学習で得た方策に基づく行動決定

4.3 手法の詳細

4.3.1 Q 学習

学習フェーズでは，Q 値の更新量が閾値を下回り，学習が収束するまで，以下の (1) から (5) の手順を繰り返す．

(1) センサによる計測

ロボットに取り付けたセンサにより環境内を計測する（ここでは，測距センサを想定する）．センサの計測範囲内に歩行者が入ると，ロボットから歩行者までの距離 l_p と角

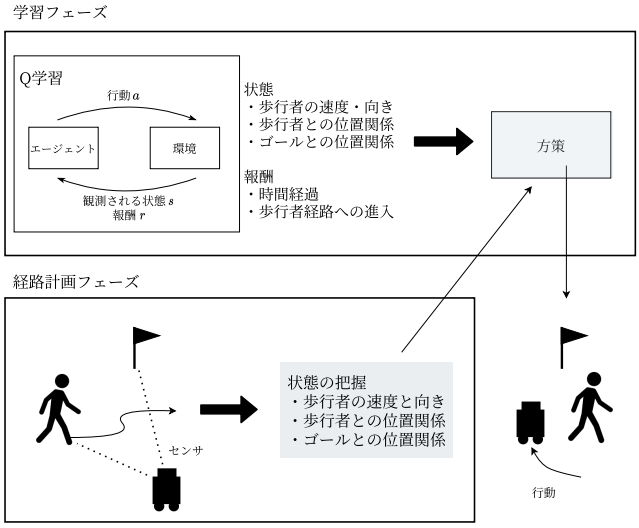


図 1: 経路計画の概要．方策は Q 学習により事前に生成しておく．方策と自律移動ロボットが計測したセンシングデータを用いて，歩行者との衝突回避及び目標地点到達のための行動を獲得する．

度 ϕ_p が計測され，計測値がエージェントに渡される．歩行者が計測範囲内に存在しない場合は，計測されなかったことを計測結果とする．

(2) 状態の更新

計測値に基づいて状態の更新を行う．環境の状態を表現する要素として，l_p と ϕ_p の他，歩行者の速度 v_p，歩行者の向き θ_p，ゴールまでの距離 l_g と角度 ϕ_g を用いる．これらの値は全てロボットに固定されたローカル座標系における値とする．上記 6 つの要素の値の組み合わせを 1 つの状態とする．

上述のとおり，センサから直接計測される値は l_p と ϕ_p である．歩行者の位置座標は，自己位置推定に基づくロボットの姿勢から算出可能であり，ゴールの位置座標は既知である．そのため，v_p，θ_p については，歩行者の位置座標の時間変化を計測周期で割ることにより算出可能であり，l_g，ϕ_g については，ロボットの自己位置座標より算出可能である．以上をまとめると，状態ベクトル

$$\boldsymbol{x} = (l_p,\ \phi_p,\ v_p,\ \theta_p,\ l_g,\ \phi_g) \tag{5}$$

が定まる．

(3) ε-グリーディ方策に基づく行動決定

状態に対応する Q 値を参照するにあたり，連続空間における状態を離散化することを考える．連続空間に属する全ての状態に対して価値を求めることは不可能なため，状態空間の各次元をそれぞれ幅 $w_{l_p}, w_{\phi_p}, w_{v_p}, w_{\theta_p}, w_{l_g}, w_{\phi_g}$ によりグリッド化し，各離散状態 $s \in \mathcal{S}$ に対して価値を求める．各グリッドに番号を振り，これが $\boldsymbol{i} = (i_{l_p},\ i_{\phi_p},\ i_{v_p},\ i_{\theta_p},\ i_{l_g},\ i_{\phi_g})$ である離散状態を $s_{\boldsymbol{i}}$ と表すことにする．

そして，(2) で算出した現在の状態 $\boldsymbol{x}$ から，離散空間における番号の組 $\boldsymbol{i}$ を求める．各離散状態において各行動 $a \in \mathcal{A}$ の Q 値が記録されているものとし，ε-グリーディ方策に基づいて $s_{\boldsymbol{i}}$ における行動 a を選択する．

(4) 報酬の獲得

本稿では，ロボットが歩行者との衝突を避けつつ，より短い時間で目標地点に到達する行動を獲得することを目的としている．そこで，時間の経過と歩行経路への侵入に対して負の報酬を設定する．時間の経過に対する報酬 r_t は，状態遷移にかかる時間を負の報酬とし，

$$r_t(s, a, s') = -\Delta t \tag{6}$$

とする．

歩行経路侵入に対する報酬 r_c については，現時刻におけるロボットの位置と，1 から n ステップ先までの歩行者位置について衝突判定を行う．ここでは，歩行者の位置を中心として，ロボットの半径 r_{robot} と歩行者の半径 r_{ped} の和の 2 倍を一辺とする正方形領域を衝突範囲とし，ロボットが衝突範囲内に位置する場合を衝突とみなす．衝突と判定された歩行者位置が未来の時刻のものであればあるほど衝突危険度は低くなるという考えの下，$i \in \{1, 2, \cdots, n\}$ ステップ先の歩行者と衝突した際の報酬を

$$r_c(s, a, s') = \beta^{i-1} \qquad (0 < \beta < 1) \tag{7}$$

と設定する．

報酬モデル R は，時間経過と歩行経路侵入に対する報酬を足し合わせたものとして，

$$R(s, a, s') = r_t(s, a, s') + c \cdot r_c(s, a, s') \tag{8}$$

と定義する．ただし，c を歩行経路侵入に対するペナルティーの大きさを決める係数とした．この式を用いて，状態遷移により得られる報酬 R を求める．

(5) 状態 s と行動 a に対する Q 値の更新

行動により得られた報酬 R と遷移先の状態 s' を用いて，1 ステップ前の状態 s と行動 a の組の価値 $Q(s, a)$ を (4) により更新する．ただし，終端状態である場合は，$\max_{a'} Q(s', a')$ の代わりに終端状態の価値を使用する．終端状態はロボットが目標地点に到達した場合，およびロボットと歩行者が衝突した場合とし，それぞれに終端価値を設定しておく．

4.3.2 経路計画

経路計画フェーズでは，学習フェーズで得られた方策を基に，毎時刻の行動を選択する．具体的には，目標地点に到達するまで以下の (1) から (3) の手順を繰り返す．経路計画フェーズでは，行動価値関数の更新は行わない．

(1) センサによる観測

前節と同様の方法で，l_p, ϕ_p を計測する．

(2) 状態の更新

前節と同様の方法で $l_p,\ \phi_p,\ v_p,\ \theta_p,\ l_g,\ \phi_g$ を算出し，現在の状態ベクトル $\boldsymbol{x} = (l_p, \phi_p, v_p, \theta_p, l_g, \phi_g)$ を得る．

(3) 学習結果に基づく行動決定

歩行者を観測している場合は，前節で算出した現在の状態 $\boldsymbol{x}$ に対する離散状態空間におけるインデックス $\boldsymbol{i}$ を求める．そして，前節で獲得した行動価値関数を参照し，s_i における行動をグリーディ方策により選択する．

5. 評価実験

本章では，提案手法による行動計画の有用性を検証するため，シミュレータ上の環境で経路計画実験を行い，従来の経路計画手法であるポテンシャル法と安全性および効率性の比較を行う．また，学習の状態に，目標位置までの距離 l_g を含める場合と含めない場合の行動計画を比較し，状態の違いが学習結果に及ぼす影響について考察する．

5.1 実験の設定

5.1.1 シミュレーション環境

学習，経路計画ともに，Python で実装したシミュレータを用いて評価を行った．シミュレーション環境は，10 [m] × 10 [m] の正方形領域内に，正方形の中心を原点とする二次元直行座標系を設定し，環境内に 1 台のロボット，1 つのゴール，1 人の歩行者を配置した．ロボットの半径は $r_{rob} = 0.2$ [m]，歩行者の半径は，歩行者自身と個人空間を考慮して $r_{ped} = 0.5$ [m] とした．歩行者は，シミュレータ外部から与える歩行者移動軌跡データを使用して移動し，ロボットはセンサによりこれを観測できるものとする．

なお，ロボットに搭載するセンサは，HOKUYO 製の測域センサ URG-04LX-UG01 の利用を想定し，製品仕様を参考に観測範囲および観測誤差を設定した．観測可能距離は 0.5 ∼ 4.0 [m]，観測可能角度は，ロボット正面を 0 [rad] として $-2\pi/3 \sim 2\pi/3$ [rad] とした．ロボットの行動 $a \in \mathcal{A}$ は，ロボットへの制御指令である前方方向への速度 ν [m/s] と中心の角速度 ω [rad/s] の組 $(\nu, \omega)^\mathsf{T}$ をそのまま用いた．ロボットの選択できる行動の集合 $\mathcal{A}$ は $\mathcal{A} = \{$ 左回転, 直進, 右回転 $\} = \{(0.0, 2.0), (1.0, 0.0), (0.0, -2.0)\}$ とした．

離散状態の設定は，センサの観測範囲に基づき決定される各状態の最小値，最大値と，離散幅により決定した．離散化のパラメータを **表 1** に示す．

5.1.2 歩行者移動軌跡データ

本稿では，博物館や大型商業施設のように道幅が一定程度あり，通路内を歩行者が行き来するような環境を想定している． そこで，想定と近い環境の人移動軌跡データセットとして ETH Dataset [14] および UCY Dataset [15] を用いることにした．これらは，市街地の歩行者を鳥瞰視点で撮影したシーンからなるデータセットであり，歩行者 4，5 人程度が並んで歩けるほどの道幅で双方向に移動可能な通路における人移動軌跡データから構成されている．UCY Dataset のシーン例を **図 2** に示す．本稿における実験では，動画像から各歩行者の軌跡を抽出し，位置座標系列に変換された後の移動軌跡データを利用した．具体的には，学習には ETH Dataset を使用し，経路計画実験には UCY Dataset を使用している．

表 1: 離散化のパラメータ

	l_p	ϕ_p	v_p	θ_p	l_g	ϕ_g
最小値	0.5	$-2\pi/3$	0.5	$-\pi$	0.0	$-2\pi/3$
最大値	4.0	$2\pi/3$	2.5	π	4.0	$2\pi/3$
離散幅	0.5	$\pi/6$	1.0	$\pi/6$	0.5	$\pi/6$

図 2: UCY Dataset のシーン例 [15]．市街地の歩行者を鳥瞰視点で撮影している．歩行者 4，5 人程度が並んで歩けるような道幅を持つ双方向に移動可能な通路における人移動軌跡データから構成されている．

学習用データセットは以下のようにして作成した．データセットから，移動距離の短い歩行者や停留している歩行者の軌跡を除外し，残った歩行者データをシミュレーション環境の領域を通過するように平行移動させた．シミュレーション環境の時間軸の離散幅は 0.1 s であるが，ETH Dataset は 0.4 s 間隔で記録された位置座標であるため，各フレーム間では歩行者は等速直線運動しているものと仮定してデータを補間した．また，学習のためのデータ数を増やすため，正規化後の各データの位置座標を原点に関して対称移動させたデータを作成し，データセットに追加した．

5.1.3 強化学習による学習処理

シミュレーション環境において，4 章で説明した手順に従って学習を行った．シミュレータにおける学習の例を **図 3** に示す．提案手法では，環境内の歩行者は 1 人であるとしているが，学習時は歩行者を観測する機会を増やすために同時に 3 人の歩行者を環境内に配置した．ただし，最初に観測した歩行者のみを観測可能とし，その歩行者の移動が終了するまで状態の更新および衝突判定にその歩行者の情報を用いることで，環境内の歩行者が 1 人である場合と同様に学習できるようにした．以下のいずれかの場合はタスク終了と判定し，目標位置，ロボットの姿勢，歩行者軌跡などをリセットする．

- ロボットが目標地点に到達した場合

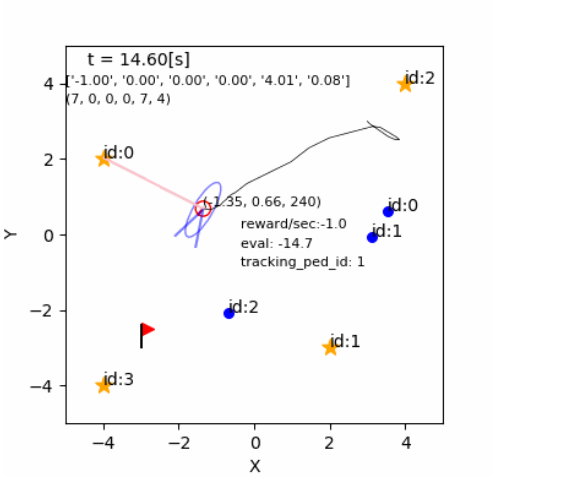


図 3: シミュレータによる学習の例．赤丸はロボット，青丸は歩行者，旗は目標地点（ゴール）である．星印は自己位置推定用のランドマークであり，実験結果に直接影響を及ぼすものではない．歩行者とランドマークには識別用の ID を割り当てている．黒線はロボットの移動軌跡を表し，ピンク色の線分はセンサによる観測を意味する．青色の楕円はカルマンフィルタによる自己位置推定における信念分布を表し，ロボットの右上に推定姿勢 (x, y, θ) が表示されている．$reward/sec$ は時刻 t に獲得した報酬，$eval$ は時刻 t までの累計報酬を表す．

- ロボットが歩行者と衝突した場合
- 試行開始から 30 秒経過した場合

リセット時，目標位置，ロボットの初期姿勢（位置・向き）はランダムに設定する．歩行者軌跡データは，移動終了時およびリセット時にデータセットの中からランダムに選択した．報酬モデルにおけるパラメータは $n = 20$，$\beta = 0.8$，$c = 200$ とした．本稿では歩行者回避行動の獲得が主な目的であるため，Q 値の初期値は目標位置に向かう行動の価値が他の行動の価値よりも高くなるように設定し，目標位置へ向かう行動を取りやすくすることにより学習にかかる時間を短縮した．学習の状態に目標位置までの距離 l_g を含めない場合を手法 1，含める場合を手法 2 とし，手法ごとに 35 万秒分の学習を行った．

5.1.4 比較手法の設定

比較手法として，ポテンシャル法を実装した．ポテンシャル法によるロボットの行動は，障害物と目標位置にそれぞれポテンシャル関数 P_o，P_g を定義し，P_o，P_g の重ね合わせによるポテンシャル場 P の勾配を求めることで決定される．本実験におけるポテンシャル場の計算式は，(x_r, y_r) をロボットの位置座標，w_o, w_g をそれぞれ障害物と目標位置のポテンシャル関数に対する重みとしたとき，

$$P(x_r, y_r) = \sum w_o P_o + w_g P_g \qquad (9)$$

と計算することにした．なお，$w_o = 0.1$, $w_g = 3.0$ とした．

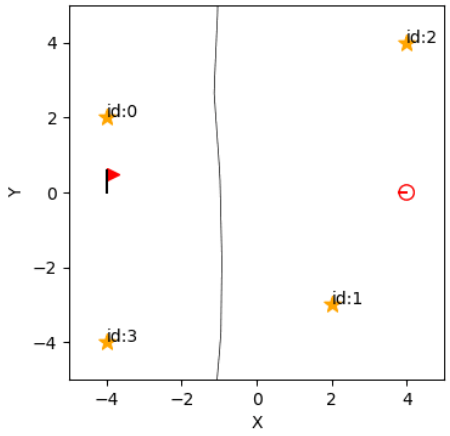


図 4: 環境の設定（セット 1 の例）．赤丸はロボットの初期姿勢，黒線は歩行者軌跡，旗は目標位置を表す．

5.2 実験の手順

ロボットが目標位置に向かう間に歩行者と遭遇するようにロボットの初期姿勢，目標位置，歩行者を設定し，前章の手順に従って経路計画を行う．ただし，歩行者を観測していない場合はロボットの向きと目標位置への角度によって，左回転，直進，右回転から行動を選択することとする．歩行者と衝突した場合，もしくは実験開始から 30 秒経過しても目標地点に到達できなかった場合を失敗とする．ここでは，ロボットの初期姿勢，目標位置，歩行者の組み合わせを 5 セット用意し，セット毎に 30 回の試行を行った．評価尺度としては，1 セットごとの成功率，ゴール到達までにかかった平均時間（ANT : Avarage Navigation Time），歩行者との最小距離の平均（AMD : Average Minimum Distance）を用いる．ポテンシャル法により行動するロボットについても同様に実験を行い，結果を比較した．

表 2 に実験で使用した設定を，**図 4** にセット 1 の設定例を示す．セット 1 からセット 4 はロボットと目標位置の間を歩行者が横切るような設定である．このうち，セット 1 とセット 3 の歩行者はロボットから一定の距離を保った位置を移動し，セット 2 の歩行者はロボットの前方から，セット 4 の歩行者はロボットの後方から接近する．セット 5 では，ロボットの前方にある目標位置の方向から歩行者がロボットに向かって接近する．

表 2: 経路計画実験の設定．ロボット位置は初期姿勢 (x, y, θ)，目標位置は位置座標 (x, y)，歩行者は歩行者軌跡データの Y 軸に対する移動方向を示す．

セット	ロボット位置	目標位置	歩行者
1	$(4, 0, -\pi)$	$(-4, 0)$	正
2	$(3, 3, 0)$	$(-2.8, -2.8)$	正
3	$(4, 0, -\pi)$	$(-4, 0)$	負
4	$(3, 3, 0)$	$(-2.8, -2.8)$	負
5	$(0, 4, -\pi/2)$	$(0, -4)$	正

5.3 実験の結果

提案手法とポテンシャル法の比較結果を **表 3** から **表 7** に示す．セット 1，3，4 では手法 1，2 ともにポテンシャル法に比べて目標位置への到達時間は長くなるが，成功率は同程度になった．セット 4 の提案手法における歩行者との距離は，ポテンシャル法に比べて 3 倍ほど長くなった．セット 2 では手法 2 の成功率が約 8 割と他の手法に比べて低く，歩行者との最小距離も短くなった．セット 5 の提案手法における成功率は非常に低く，手法 1 においては歩行者との衝突を回避できず，目標位置に到達できなかった．

6. 考察

6.1 状態空間の差

手法 1 と手法 2 の違いについては，ほとんどの場合の成功率は同程度だが，手法 1 は手法 2 に比べて目標位置への到達時間，歩行者との距離がともに短くなっている．手法 1 の歩行者との最小距離は，$r_{rob} = 0.2$ [m]，$r_{ped} = 0.5$ [m] と設定していることから十分な距離を取っていると考えられ，安全な衝突回避を行なっているといえる．セット 2 における手法 2 の成功率は手法 1 に比べて低くなっているが，これは歩行者が観測範囲に入った際，手法 1 では方向転換して歩行者の進行方向に進まない行動を取り，歩行者と接近した際には歩行者の後方を通ろうとしているが，手法 2 では歩行者が接近するまで目標位置に向かう行動を取り続けるため，衝突する可能性が高まるためと考えられる．

この行動の違いの要因としては，手法 2 は学習の状態に l_g を含めることで離散状態数が手法 1 の 8 倍になっていることが考えられる．そのため，同じ学習時間では学習が不十分な可能性がある．学習が不足している場合には，学習時に設定した行動価値関数の傾向が残っており，衝突回避より目標位置に向かう行動を優先したと考えられる．

以上より，以降，手法 1 とポテンシャル法の比較を行う．

表 3: 実験結果（セット 1）

手法	成功率	ANT[s]	AMD[m]
手法 1	0.97	11.22	1.35
手法 2	0.97	11.44	1.42
ポテンシャル法	1.00	12.11	2.19

表 4: 実験結果（セット 2）

手法	成功率	ANT[s]	AMD[m]
手法 1	1.00	11.20	1.00
手法 2	0.77	11.28	0.94
ポテンシャル法	1.00	10.66	1.40

表 5: 実験結果（セット 3）

手法	成功率	ANT[s]	AMD[m]
手法 1	1.00	11.07	1.55
手法 2	1.00	11.29	1.63
ポテンシャル法	1.00	11.45	2.18

表 6: 実験結果（セット 4）

手法	成功率	ANT[s]	AMD[m]
手法 1	1.00	12.43	2.81
手法 2	1.00	12.52	2.85
ポテンシャル法	0.97	11.77	0.96

表 7: 実験結果（セット 5）

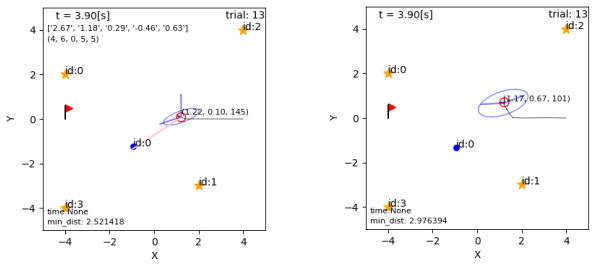
手法	成功率	ANT[s]	AMD[m]
手法 1	0.00	-	-
手法 2	0.03	11.70	0.95
ポテンシャル法	0.70	10.09	1.15

6.2 安全性について

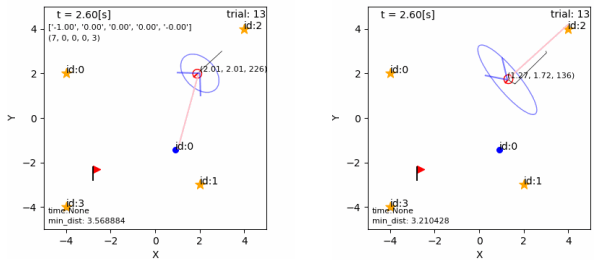
セット 1 からセット 4 のようにロボットと目標位置の間を歩行者が横切る場合には，両手法の成功率は同程度であった．歩行者との最小距離の平均を比較すると，ポテンシャル法より 0.4 ～ 1.9 [m] 短いが，ロボットの半径 r_{rob} と個人空間を含めて設定した歩行者の半径 r_{ped} の和である 0.7 [m] 以上となっており，歩行者から十分な距離を取って行動しているといえる．そこで，回避行動の軌跡を比較してみることにした．手法 1 とポテンシャル法の軌跡の違いを **図 5** に示す．ポテンシャル法によるロボットは，歩行者が観測範囲に入ると歩行者と並行に同方向への進路を取り，歩行者が目前に迫る（または目標位置から遠ざかる行動を取る直前になる）と，その場で回転することで衝突を回避している．一方，手法 1 によるロボットは歩行者が観測範囲に入ると右回転，左回転を繰り返しながら少しずつ直進し，歩行者が自身の正面を通過するのを待ってから目標位置へ向かうことで衝突を回避している．

一方，セット 5 においては，手法 1 により衝突を回避できていない．経路計画が失敗した原因としては以下の 2 点が考えられる．まず，歩行者を避けるまでに時間がかかることである．セット 5 では歩行者がロボットの前方から正面に向かって接近し，短時間で両者の距離が縮まる．手法 1 では歩行者の通過を待ってから行動する傾向があり，このような状況下では回避行動を取るまでに与えられた時間が短くなるため衝突してしまう．これらは，歩行経路侵入に対する負の報酬を大きくする（歩行経路内にとどまる行動の価値を下げる）ことや，行動の選択肢を増やす（$\mathcal{A}$ に速い直進や後方へ下がる行動などを加える）ことで解決できると考えらる．もう 1 点は，センサ値や算出した歩行者の速度，向きにおける誤差である．**図 6** に真の状態と観測された状態 $\boldsymbol{x}$ が異なっている例を示す．このような誤差により，真の状態と異なる状態における方策を参照し，状況に適した行動をとっていない可能性がある．これらの解決策としては，観測精度の高いセンサの使用や離散幅を調

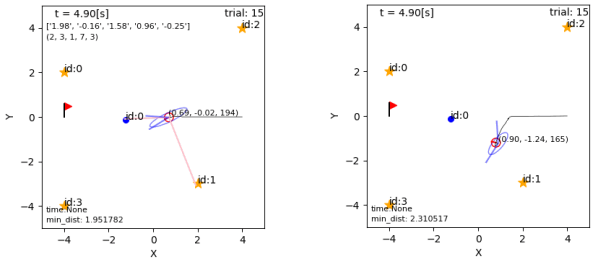
Vol.2021-DPS-186 No.63
Vol.2021-CSEC-92 No.63
2021/3/16



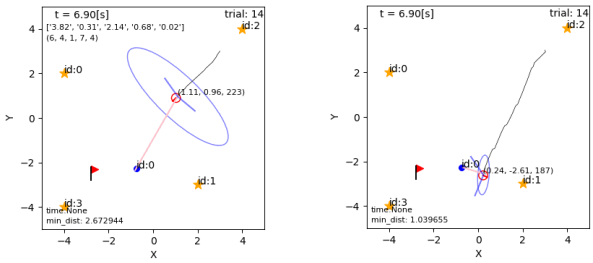
(a) セット 1



(b) セット 2



(c) セット 3



(d) セット 4

図 5: 歩行者が観測範囲に入った際の行動による軌跡の違い．左に手法 1，右にポテンシャル法の軌跡を示す．30 回試行した中から代表的なものを選んだ．(a) から (c) のポテンシャル法においては，ロボットが歩行者の進路と並行に同方向に進んでいる．(d) では移動した後に目標位置の延長線上で回転し，目標位置から遠ざからない行動かつ歩行者と衝突しない行動をとっている．手法 1 では (a) から (d) すべてにおいて歩行者から離れた位置にいる．

整することで誤差を小さくする方法や，歩行者の行動予測を取り入れることが考えられる．歩行者を見失った場合や観測値がない場合に予測値で補填する，1 ステップ前との観測値の差が大きい場合には予測値に置き換えて状態を算出するなどの改善が見込める．

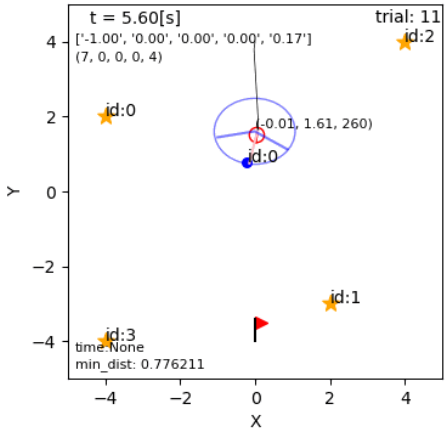


図 6: セット 5 における手法 1 の観測誤差の例．$t = 5.60$ [s] におけるロボットと歩行者の距離は人との最小距離と等しく 0.78 [m] であったが，観測値 l_p は，歩行者を観測していないことを意味する -1.00 であった．

6.3 効率について

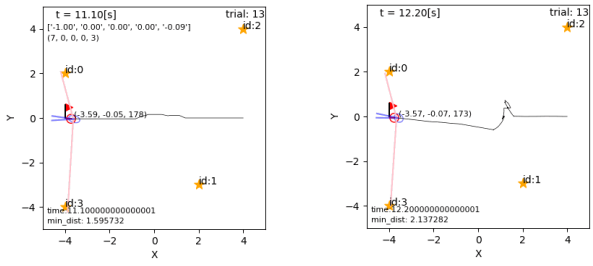
目標位置到達にかかる時間の平均は，全てのセットにおいて手法 1 がポテンシャル法より遅くなった．しかし，**図 7** に示すゴール到達までの軌跡を見ると，ポテンシャル法では進路を引き返す行動やその場で複数回回転する行動，大きく迂回する経路などが見られる．手法 1 では初期位置から目標位置までを直線的に結ぶ経路を進んでいることから，ポテンシャル法の方が効率的であるとも言えない．

手法 1 の目標位置への到達時間が遅い要因としては，ロボットの取り得る行動が限定されており，歩行者が通過するまでその場で回転する場合が多いことが考えられる．また，Q 学習により獲得した方策も一因に挙げられる．手法 1 では歩行者の前方を通ると衝突可能性が高まることから，歩行者が正面を通過するまで前進しない行動を学習した可能性がある．観測範囲内の歩行者の有無や歩行者との距離によって制御指令値 $(\nu, \omega)^\mathsf{T}$ を変える機能を追加することや，歩行者の速度によって歩行者進路への侵入に対する報酬を変化させることで，歩行者の接近速度が遅い場合には通過を待たずに行動するなどの改善が期待できる．

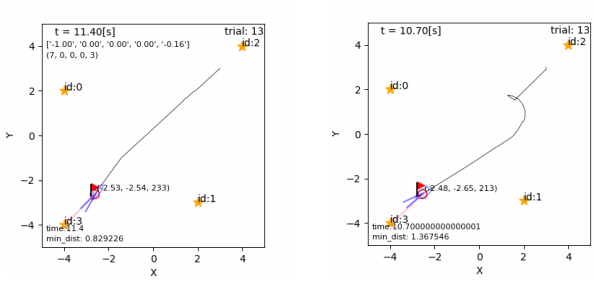
7. おわりに

本稿では，人・ロボット共存環境における安全かつ効率的な経路計画を目的に，ロボットと歩行者，目標地点の位置関係および歩行者の移動傾向を状態とする強化学習による経路計画手法を提案した．提案手法による経路計画実験では，歩行者がロボットと目標位置の間を横切るような多くの場合において，ポテンシャル法と同等の成功率で歩行者回避を実現できることがわかった．また，目標位置到達にかかる時間は多少長くなるもののロボットは歩行者の進路を妨げないように通過するのを待ち，歩行者の後方を通行するといった行動を学習していることも確認できた．

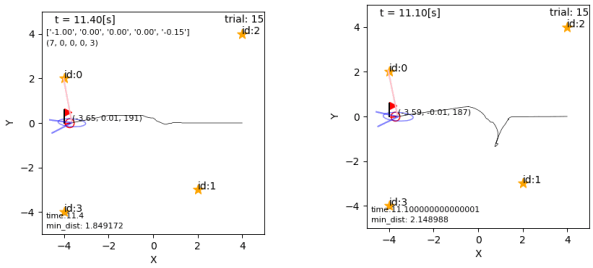
Vol.2021-DPS-186 No.63
Vol.2021-CSEC-92 No.63
2021/3/16



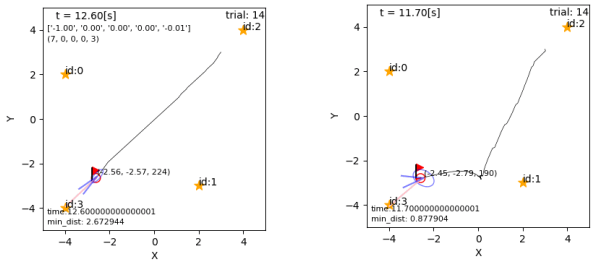
(a) セット 1



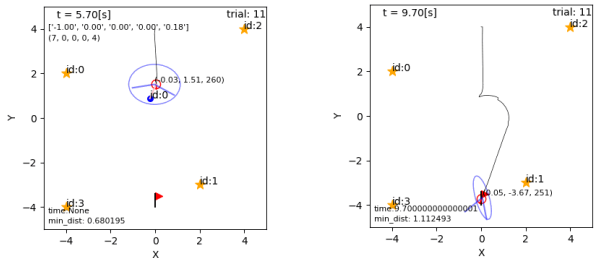
(b) セット 2



(c) セット 3



(d) セット 4



(e) セット 5

図 7: 各セットにおける手法 1 によるロボットの軌跡（左列）とポテンシャル法による軌跡（右列）．各 30 回試行した結果から代表的なものを 1 つ示す．(e) の手法 1 における軌跡は衝突時のものである．

一方，歩行者との距離が近くなってからロボットが歩行者を発見した場合や歩行者がロボットに向かって接近するような状況では，歩行者を回避しきれないという課題も明らかになった．今後，歩行者進路への侵入に対する報酬を変動させることやロボットが選択可能な行動を追加すること，状況に応じてポテンシャル法への切り替えを行うことで本課題に対応していきたい．また，本稿では，観測範囲内に存在する歩行者は一人であることを仮定したが，今後，歩行者が複数存在する状況にも対応していく予定である．

謝辞 本研究の一部は，電気通信普及財団，JSPS 科研費 20K11776，20K12011 の助成を受けたものである．

参考文献

[1] 彌城祐亮，江口和樹，岩崎 聡：ポテンシャル法によるロボット製品の障害物回避技術の開発，三菱重工技報，Vol. 51, No. 1, pp. 40–45 (2014).
[2] 平瀬祐貴，三輪顕太朗，山内悠嗣：動的環境を考慮した移動ロボットの経路計画，*ROBOMEC2019*, pp. 1A1–F10 (2019).
[3] 野口博史，山田隆基，森 武俊，佐藤知正：大量の人移動計測データに基づく移動ロボットの人回避経路計画，日本ロボット学会誌，Vol. 30, No. 7, pp. 684–694 (2012).
[4] 縄田 翔，桜間一徳，中野和司：動的障害物を考慮したナビゲーション関数による衝突回避制御，自動制御連合講演会，p. 142 (2010).
[5] 岩朝睦美，戸田雄一郎，久保田直行：予測可能な移動障害物のある環境における時空間グラフを用いた大域的経路探索と行動計画，日本機械学会論文集，Vol. 85, No. 876, pp. 18–00254 (2019).
[6] Todi, V., Sengupta, G. and Bhattacharya, S.: Probabilistic Path Planning using Obstacle Trajectory Prediction, *Proc. of ACM CODS-COMAD 2019*, pp. 36–43 (2010).
[7] Sutton, R. S. and Barto, A. G.: *Reinforcement Learning: An Introduction*, The MIT Press (2018).
[8] 五十嵐治一：強化学習を用いた自律移動型ロボットの行動計画法の提案，人工知能学会論文誌，Vol. 16, No. 6, pp. 501–509 (2001).
[9] 平岡賢治，青柳誠司：強化学習を用いた移動ロボットの障害物回避軌道の探索：A*アルゴリズムと強化学習の統合，*ROBOMEC2008*, pp. 2P2–G05 (2008).
[10] Jing, Y., Chen, Y., Jiao, M., Huand, J., Niu, B. and Zheng, W.: Mobile Robot Path Planning Based on Improved Reinforcement Learning Optimization, *Proc. of the 2019 International Conference on Robotics Systems and Vehicle Technology*, pp. 138–143 (2019).
[11] 有馬純平，黒田洋司：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，人工知能学会全国大会，pp. 4Rin1–19 (2019).
[12] 武田真人，長尾智晴：移動障害物回避を実現する予測型強化学習の提案，電気学会論文誌 C，Vol. 129, No. 6, pp. 1115–1122 (2009).
[13] Watkins, C. J. and Dayan, P.: Q-learning, *Machine Learning*, Vol. 8, No. 3–4, pp. 279–292 (1992).
[14] Pellegrini, S., Ess, A., Schindler, K. and Cool, L.: You'll Never Walk Alone: Modeling Social Behavior for Multi-target Tracking, *Proc. of ICCV 2009*, pp. 261–268 (2009).
[15] Lerner, A., Chrysanthou, Y. and Lischinski, D.: Crowds by Example, *Computer Graphics Forum*, Vol. 26, No. 3, pp. 655–664 (2007).