Vol.2019-ICS-194 No.2
2019/3/9

巡回問題における能力の異なる複数エージェントの自律的な行動決定手法

岩田 裕登[1,a)]　杉山 歩未[1,b)]　菅原 俊治[1,c)]

概要：本研究では，マルチエージェント協調巡回問題において不均一な移動速度のエージェントが直接通信をせずに自らの移動速度に応じて分業を実現する手法を提案する．近年，ロボット技術は急速に発展しており，活動分野が大きく拡大している．ロボット 1 台のみでは対応できない作業範囲では，複数台で協調しながら作業することが期待されている．複数台での作業では各エージェントの能力が異なる場合もあり，自律的に自らの能力に適した分業を実現するメカニズムは明らかになっていない．本研究では，移動速度が異なるエージェントを用いた巡回清掃において，自らの移動速度に応じて分業する手法を提案する．移動速度によって選択する戦略や清掃する箇所に違いが見られた．一方，提案した手法は特定の環境においては移動速度を反映した分業が実現できることが分かった．

Autonomous Target Decision Method for Heterogeneous Multi-Agent Patrolling Task

1. はじめに

近年，ロボット技術の進歩により，ロボットの活動範囲が急速に拡大している．清掃，警備，原子力施設や宇宙空間などの危険な場所での作業でもロボットが活躍している．こうした作業の中には，広大な範囲の作業，継続的に行うことが期待される作業，短時間で遂行するべき作業などもある．1 台のロボットのみでは作業できる範囲，移動速度，バッテリ残量などが限られており，こうした作業は困難である．これらの場合，複数台のロボットが協調して作業する必要がある．

本研究では複数台での協調作業の中から，巡回清掃を題材として扱っている．複数台のロボットによる巡回清掃において，ロボットは環境全体をカバーしつつ，汚れやすい箇所はより多くの頻度で巡回する必要がある．また，ロボットにはバッテリがあり，定期的に充電基地へ戻る必要がある．さらに，複数のロボットが同じ箇所に集中しないように他者との協調や競合回避が必要である．これらの状況下で最適な巡回を実現する手法は自明でない．このような課題を解決するために，ロボットの制御プログラムをエージェントとして捉え，エージェント間の協調により効率的な作業を実現するというアプローチがある．エージェントは環境の情報，他のエージェントの情報，自らの情報を考慮して，適切な行動戦略を選択することが求められる．

複数のエージェントが高度に協調する手段として，エージェント間で直接通信して多様な情報を交換することが考えられる．しかし，情報の交換と受け取った情報の処理には大きな計算量が必要である．清掃や警備で使用されるロボットは計算機能，バッテリ容量に制限がある．そこで，我々は双方の送受信を必要とする直接的な通信を制限しながら，できる限り協調して高い効率を出すことに着目した．

複数エージェントの協調に関する研究の多くでは全エージェントの特性が均一であると仮定して実験している．しかし，エージェントが常に同一の特性をもつとは限らない．例えば清掃ロボットにおいては，新しい機能が追加される，動作の効率が改善されるなどの技術革新により，後の時期に製造されたものが，以前に製造されたものよりバッテリの持ちがよく，速く移動できることが考えられる．特性が不均一である状況でエージェント間で通信し，移動速度などの特性を考慮して，担当領域をそれぞれのエージェント

[1] 早稲田大学
Waseda University, Tokyo, 169–8555 Japan
[a)] y.iwata@isl.cs.waseda.ac.jp
[b)] a.sugiyama@isl.cs.waseda.ac.jp
[c)] sugawara@waseda.jp

に割り当てるアプローチが存在する．しかし，直接の通信が制限された状況で特性に適した協調を実現する手法は明らかになっていない．

本研究では，エージェントの特性のうち移動速度が不均一な状況で，エージェント間で直接通信せずに移動速度に適した協調を実現する手法を提案し，その有効性を実験的に調査する．

2. 関連研究

単数または複数のエージェントでの巡回問題に関する研究は数多く存在する．[1] では連続巡回問題を定式化して，イベントの発生頻度が地点により異なる環境において 1 台のエージェントが地点ごとに異なる頻度で巡回する手法が提案された．しかし，ここでは複数台での協調作業は考慮されていない．[2] は [1] を拡張して，複数台での協調巡回にも対応した．複数台のロボットが通信によって情報を交換することで担当領域を分割している．ロボットの速さが不均一な場合は速さに合った広さの領域に分割して適応する．[3] では，全体をカバーする経路を生成し，それをコストが最小になるように分割するオフラインでの協調巡回手法が提案された．

一方，領域を分割せずに他エージェントの行動や環境の情報に基づいて適切な戦略をとって巡回する手法が存在する．例えば，[4] では，エージェント間の直接の通信が制限された状況で，各エージェントがいくつかの戦略の中からごみの存在期待値や各戦略で回収したごみの量に基づいて適切な戦略を自律的に学習して選択する手法が提案されている．[5] ではエージェントが通信できる範囲とメモリに制限を設け，通信にノイズがある状況で，分散して巡回する手法が提案された．

ここまでに示した研究の多くではエージェントの特性が均一であると仮定しているが，特性が異なるエージェントを用いた研究も存在する．[6] では地上の一定領域を視野に持つ複数の飛行ロボットによる地上の巡視を題材とし，地上全体をカバーできる閉路を生成してエージェントごとに分割して巡視する．隣接エージェント間の通信で互いの担当領域・移動速度の情報を送受信して，情報をもとに互いの担当領域を再計算する．しかし，隣接エージェントのみと交渉することで領域を分割する手法は収束に時間がかかる．また，[4] のような通信がより制限された状況で自律的に特性に適した分業を促すメカニズムは明らかになっていない．本研究では，[4] のモデルをもとに，能力が異なる状況として移動速度が不均一な場合を想定し，自律的に特性に適した分業を実現する手法を提案する．

3. モデル

3.1 環境の仮定

本論文では [4] で使用されたモデルを拡張している．そのため，[4] で用いられた仮定について述べる．まず，エージェント間での直接の通信は行わない．すなわち，他のエージェントの経路や戦略などの情報を取得できない．これは，エージェント間で通信する情報を制限することで処理が減るため，また，異なるプロトコルを持つなどの理由で通信ができないエージェントとの協調を考慮するためである．次に，エージェントは自分および他のエージェントの位置，および移動速度が分かるとする．自分と他のエージェントの位置が分かる環境の一つに，環境上に赤外線を発する装置，各エージェントに赤外線を反射する素子があり，環境上の装置が反射する赤外線からエージェントの位置を特定し，その位置情報を全エージェントにブロードキャストする環境が想定される．また，エージェントの移動速度は時間による位置の変化から計算することができる．また，エージェントは環境の構造および環境のごみの発生確率の分布を既知と仮定した．本来，エージェントはこれらの情報をはじめから知ることはできないかもしれない．しかし，構造の情報を生成するアルゴリズムは [7] など多数存在している．また，環境の各地点のごみの発生確率を学習する手法が [8] で提案されている．本研究では，環境や自らの特性に適した戦略を決定する手法に着目するため，このような仮定を導入した．最後に，複数のエージェントが同一のノード・充電基地に存在でき，衝突回避を考慮しない．衝突回避のアルゴリズムも [9] などで考案されている．本研究にこうした手法を利用することで衝突回避を実現できると考えて，上記のように仮定した．

3.2 環境とエージェント

エージェントが清掃する環境をグラフ $G = (V, E)$ と表す．$V = \{v_1, ...v_n\}$ をノードの集合，エッジの集合を E とする．ノード上にはエージェントやごみが存在することもある．各ノードの汚れやすさを確率的に表現し，その確率をごみの発生確率と呼ぶ．各ノード v のごみの発生確率を $p(v)(0 \leq p(v) \leq 1)$ とする．エッジは 2 つのノードをつなぎ，エージェントはエッジで直接つながれたノード間を移動できる．

エージェントの集合を $A = \{1, ...n\}$ とする．エージェント i の速さを s_i とする．本研究では離散時間を導入し，1 ステップをその最小単位とする．各エージェントは 1 ステップで s_i 回エッジを通って隣接ノードへ移動し，移動先のノードにあるごみを回収する．時刻 t においてエージェント i が通過してごみを回収したノードの集合を V_t^i と表す．V_t^i のうち，エージェント i が時刻 t の間で最後に到着したノードを v_t^i とする．2 つのノード $v_i, v_j \in V$ 間の最短距離を $d(v_i, v_j)$ と表す．ここでの最短距離はグラフ上の最短経路におけるエッジの長さの和とする．また，各エージェントはバッテリを持つ．エージェント i のバッテリの性能は $(B^i_{max}, B^i_{drain}, k^i_{charge})$ と表される．B^i_{max} はバッ

テリ容量，B^i_{drain} はバッテリの消費率，B^i_{charge} はバッテリの充電率とする．時刻 t におけるエージェント i のバッテリ残量を $b^i(t)(0 \leq b^i(t) \leq B_{max})$ とする．速さに関わらず，エージェントが隣接ノードへ 1 回移動すると $b^i(t)$ は B^i_{drain} 減少する．速さ s_i のエージェントは 1 ステップで s_i 回移動することから，$b^i(t)$ は 1 ステップで $B^i_{drain}s_i$ 減少する．また，各エージェント i は充電基地 v^i_{base} から出発し，v^i_{base} で充電する．充電基地で k^i_{charge} ステップ充電すると $b^i(t)$ は 1 増加する．また，エージェントはバッテリ残量が 0 になると動けなくなるが，その前に必ず充電基地に戻る．残量がなくなる前に必ず充電基地に戻るためのアルゴリズムは後に 3.4.3 で説明する．

時刻 t におけるノード v のごみの量を $L_t(v)$ とする．各ノード，各ステップにおいて $L_t(v)$ が以下の式のように更新されることでごみの蓄積を表す

$$L_t(v) \leftarrow \begin{cases} L_{t-1}(v)+1 & (\text{確率 } p(v) \text{ でごみが発生した時}) \\ L_{t-1}(v) & (\text{それ以外}). \end{cases} \quad (1)$$

ただし，時刻 t でエージェントがノード v を訪れたときはごみが回収されて $L_t(v) = 0$ となる．エージェントは他エージェントの位置が分かるため，各ノードについて最後に清掃された時刻が分かる．時刻 t におけるノード v のごみの存在期待値 $EL_t(v)$ を以下の式で計算することができる

$$EL_t(v) = p(v) \cdot (t - t^v_{visit}). \quad (2)$$

ここで t^v_{visit} はノード v が最後に清掃された時刻である．

3.3 評価指標

巡回清掃では，環境にごみが多く残らない，または長時間残らないようにごみを回収することが求められる．そこである期間の各ノードのごみの量の総和を清掃効率の評価指標とする．この評価指標を以下の式で定義する

$$D_{t_s,t_e} = \sum_{v \in V} \sum_{t=t_s+1}^{t_c} L_t(v). \quad (3)$$

ここで t_s, t_e はそれぞれ評価対象の時間の開始時刻，終了時刻であり，$t_s < t_e$ を満たす．D_{t_s,t_e} が小さいほど，清掃効率が良いといえる．

3.4 エージェントの巡回戦略

エージェントは 2 つのステージで行動計画を立てる．1 つはノードの集合 V から清掃するべきノードを決定する目標ノードの決定ステージ，もう 1 つは目標ノードまでの経路を生成する経路生成である．[4] では目標ノードを決める手法として学習型目標決定法が提案された．本研究では，学習型目標決定法を拡張した手法で目標ノードを決定する．学習型目標決定法について説明する．

3.4.1 学習型目標決定法 (AMTDS)

[4] で提案された学習型目標決定法では後に説明する複数の戦略の中から適切な戦略を決定している．戦略 s に従って目標ノードを選択してからその目標ノードまで移動した間に回収したごみの量から報酬 u を次の式で計算する

$$u = \frac{\sum_{t \in T_{travel}} \sum_{v \in V^i_t} L_t(v)}{d_{travel}}. \quad (4)$$

ここで T_{travel} は戦略 s に基づいて目標ノードを選択してからその目標ノードに到着するまでの時間，d_{travel} はその時間にエージェントが移動した経路長を表す．報酬 u を用いて戦略 s の行動価値 $Q^i(s)$ を次の式で学習する

$$Q^i(s) \leftarrow (1-\alpha)Q^i(s) + \alpha u. \quad (5)$$

ここで $\alpha(0 < \alpha \leq 1)$ は学習率である．エージェントは得られた行動価値 $Q^i(s)$ を用いて，ε-greedy 法によって戦略を決定する．ε-greedy 法においては，確率 ε で全ての戦略からランダムに戦略を選択し，$1-\varepsilon$ の確率で行動価値 $Q^i(s)$ が最大となる戦略 s を選択する．学習型目標決定法では，各エージェントが 1 ステップあたりに回収するごみの量を最大化するように適切な戦略を選んでいる．なお，学習型目標決定法では別の戦略をエージェントに追加することで，その戦略も選択肢となる．

3.4.2 目標ノードの選択戦略

学習型目標決定法においてエージェントは戦略に応じて目標ノード v^i_{tar} を以下のように決定する．

- ランダム法: ノードの集合 V 全体からランダムに目標ノード v^i_{tar} を一つ選択する．
- 貪欲法: ノードの集合 V 全体のうち，ごみの存在期待値 $EL_t(v)$ の上位 N_g 個の集合を V^t_g とする．ただし，$EL_t(v)$ が N_g 番目のノードが複数個存在する場合は，そのすべてを V^t_g に含める．V^t_g からランダムに目標ノード v^i_{tar} を一つ選択する．V^t_g からランダムに選ぶようにしたのは，複数のエージェントが特定の上位ノードを集中して目標ノードとすることを防ぎ，目標ノードを分散させるためである．
- 斥力法: ノードの集合 V 全体のうち，ランダムに N_{rep} 個を選んで集合 V^i_{rep} とする．V^i_{rep} の中で式 (6) に示す通りに全エージェントからの距離の合計値が最も大きいものを目標ノード v^i_{tar} として選択する．

$$v^i_{tar} = \arg\max_{v \in V^i_{rep}} \sum_{i \in A} d(v^i_t, v). \quad (6)$$

- 戦略型目標決定法: 近くに汚れたノードがある場合に近いノードから目標ノードを選択する戦略である．自エージェントから一定距離 d_{rad} 内にあるノードの集合を V^i_{area} とし，V^i_{area} 内のごみの存在期待値の平均を EV^i_t とする．平均値 EV^i_t がしきい値 EV_{th} より多い場合には V^i_{area} から目標ノード v^i_{tar} を選択し，そうで

ない場合はノード全体 V から貪欲法で目標ノードを選択する．目標ノードに到着したときにしきい値 EV_{th} を以下の式で学習する

$$EV_{th} \leftarrow EV_{th} + \alpha(EV_{t+k} - EV_{th}). \quad (7)$$

ここで，α はしきい値の学習率，EV_{t+k} は目標ノードに到着した時点での自エージェントから一定距離 d_{rad} 内にあるノードの集合 V^i_{area} のごみの存在期待値の平均である．

3.4.3 経路生成

[4] では経路生成の手法として最短経路およびサブゴール型経路計画法（the *gradual path generation*, GPG）が導入された．本実験ではサブゴール型経路計画法が常に有効な経路を生成することから，本手法のみを用いることにする．

サブゴール型経路計画法の説明の前に，この手法で使用するポテンシャルについて説明する．エージェント i は各ノード v についてそのノードから充電基地 v^i_{base} に戻るために必要な最小のバッテリ残量を計算する．この最小のバッテリ残量をポテンシャルと呼び，ノード v のポテンシャル $\mathcal{P}^i(v)$ を次の式で計算する

$$\mathcal{P}^i(v) = d(v, v^i_{base}) \cdot B_{drain}. \quad (8)$$

サブゴール型経路計画法では，基本的には最短経路に従うが，その経路の近くにごみが多いノードが存在する場合にはそこに立ち寄るように経路を生成する．エージェント i の現在ノード v^i_t を 0 番目のサブゴール $v^i_{sub,0}$ とする．エージェントは $n-1$ 番目のサブゴール $v^i_{sub,n-1}$ を用いて，下記の式 (9) の条件を満たすノードのうち，最もごみの存在期待値が大きいノードを n 番目のサブゴール $v^i_{sub,n}$ とする．

$$\left.\begin{aligned} &d(v^i_{sub,n-1}, v^i_{sub,n}) \le d_{cl} \\ &d(v^i_{sub,n}, v^i_{tar}) < k_{att} \cdot d(v^i_{sub,n-1}, v^i_{tar}) \\ &d(v^i_{sub,n-1}, v^i_{sub,n}) + d(v^i_{sub,n}, v^i_{tar}) \\ &\quad \le k_{rover} \cdot d(v^i_{sub,n-1}, v^i_{tar}) \\ &\mathcal{P}^i(v^i_{tar}) + B_{drain} \cdot (d(v^i_{sub,n-1}, v^i_{sub,n}) \\ &\quad + d(v^i_{sub,n}, v^i_{tar}) \le b_t. \end{aligned}\right\} \quad (9)$$

ここで，d_{cl} は順序が隣り合うサブゴール間の最大距離を表すパラメータ，k_{att}，k_{rover} は最短経路から遠回りできる距離の程度を表すパラメータである．決定したサブゴール $v^i_{sub,n-1}$，$v^i_{sub,n}$ が $d(v^i_{sub,n-1}, v^i_{sub,n}) \le d_{cl}$，$v^i_{sub,n} = v^i_{tar}$ を満たすとき，次のサブゴールを決める動作を終了する．そして，サブゴール間を最短経路でつなぐことで現在ノードから目標ノードまでの経路を生成する．

エージェントは充電基地まで戻ることができるバッテリ残量を保つように経路を生成する必要があるが，それができない場合は，次の目標ノードを充電基地 v^i_{base} とし，充電基地までの最短経路を生成する．

3.5 提案手法

特性が異なる分業の形態のひとつとして，充電基地からの距離による分業に着目した．エージェントは定期的に充電基地に戻らなければならないため，充電基地から遠い領域ほど満充電の状態から移動して巡回できる時間が短くなる．よって，複数台のうち一部のエージェントは遠い領域を多く巡回する必要がある．我々は速く移動できるエージェントがその速度を生かして遠いノードを回り，遅くしか移動できないエージェントが近いノードを回ることで，特性に適した分業と清掃効率の向上ができるのではないかと考えた．既存手法 [4] にこのような分業を直接促す仕組みは含まれていない．そこで，報酬式 (4) に移動速度，充電基地からの距離を組み込んだ．速いエージェントは距離が遠いほど報酬を大きくし，充電基地から遠いノードへ行くような戦略を選択することを促し，遅いエージェントは距離が近いほど報酬を大きくし，充電基地から近いノードへ行くような戦略を選択することを促した．これによって距離による分業ができることを期待した．

[4] の既存手法では報酬を式 (4) で計算していた．提案手法では報酬 u を次の式で計算する．

$$u = \frac{\sum_{t \in T_{travel}} \sum_{v \in V^i_t} L_t(v)\{1 + (s_i - s_{ave})\frac{(d(v, v^i_{base}) - d_{ave})}{d_{ave}}\}}{d_{travel}}. \quad (10)$$

ここで，s_i はエージェント i の移動する速さ，s_{ave} は全エージェントの平均の速さ，d_{ave} は環境上の各ノードと充電基地間の距離の平均である．移動速度 s_i が平均 s_{ave} よりも速いエージェントは d^i_t が大きい，すなわち充電基地から遠いほど報酬 u が大きくなる．一方，移動速度 s_i が平均 s_{ave} よりも遅いエージェントは d^i_t が小さい，すなわち充電基地から近いほど報酬 u が大きくなる．本手法は自らの特性による分業を，距離による分業として実現しようとするものである．

4. 実験と考察

4.1 実験環境

清掃する環境 G は 101×101 のグリッド構造とする．各ノード v は座標をもち，(x, y) で表す．ただし，$-50 \le x, y \le 50$ である．全エージェントの充電基地を $v^i_{base} = (0, 0)$ とする．

実験は 2 つの環境で行った．環境のごみ発生確率の分布を図 1 に示す．ごみ発生確率 $p(v)$ は図 1 における赤色の領域が 10^{-3}，灰色の領域が 10^{-4} その他の白色の領域が 10^{-6} である．環境 (a) では各地点でごみが均一な確率で発

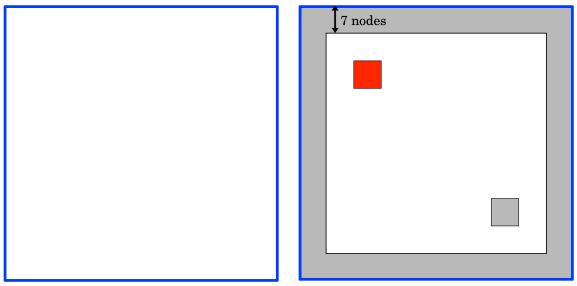

図 1 清掃する環境
Fig. 1 Cleaning environments.

表 1 目標選択戦略におけるパラメータ
Table 1 Parameters in target decision strategies.

目標選択戦略	パラメータ	値
貪欲法	N_g	5
斥力法	N_{rep}	100
戦略型目標決定法	α	0.1
	d_{rad}	15
学習型目標決定法	α	0.1
	ε	0.05

表 2 サブゴール型経路計画法におけるパラメータ
Table 2 Parameters in GPG method.

パラメータ	値
d_{cl}	15
k_{att}	1.0
k_{rover}	1.2

生する．環境 (b) ではグリッドの周囲および内側の一部の正方形領域にごみがたまりやすい．

実験において 1 回の試行のステップ数は 3000000 ステップとした．実験結果は 10 回の試行の平均値を使用した．環境には 10 台のエージェントを配置した．本研究では移動速度を不均一にした状況で実験するため，10 台のうち 5 台のエージェントの移動速度 s_i を 2，残り 5 台のエージェントの移動速度 s_i を 1 とした．また，移動速度 2 のエージェントのバッテリ消費率は 1，移動速度 1 のエージェントのバッテリ消費率は 2 とした．バッテリの性能など，移動速度とバッテリの消費率以外の特性は全てのエージェントで同一である．エージェントは既存手法である学習型目標決定法，または提案手法のいずれかによって戦略を決定し，目標ノードを選択する．目標ノードの選択戦略で使用する各パラメータの値を表 1 に示す．また，サブゴール型経路計画法で使用する各パラメータの値を表 2 に示す．

バッテリの性能を表 3 に示す．これにより，エージェントの移動速度が 1 の場合，2 の場合ともに連続で 450 ステップ稼働する．これは，エージェントの移動速度が 1 の場合，消費率 B^i_{drain} が 2 であるため，バッテリを 1 ステップあたり 2 消費し，移動速度が 2 の場合，1 ステップで 2

表 3 バッテリの性能
Table 3 Parameters in battery.

パラメータ	移動速度 1	移動速度 2
移動速度 s_i	1	2
容量 B^i_{max}	900	900
消費率 B^i_{drain}	2	1
充電率 k^i_{charge}	3	3

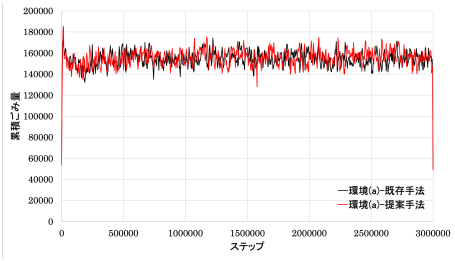

図 2 環境 (a) における既存手法と提案手法の清掃効率推移.
Fig. 2 Transition of cumulative dust accumulation over time in Env. (a).

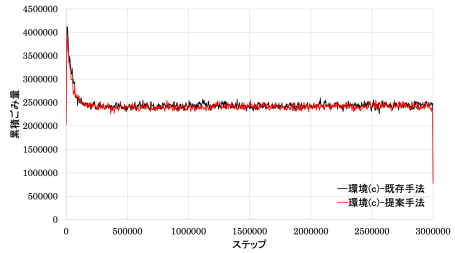

図 3 環境 (b) における既存手法と提案手法の清掃効率推移.
Fig. 3 Transition of cumulative dust accumulation over time in Env. (b).

ノード移動するため，バッテリを 1 ステップあたり 2 消費するからである．また，移動速度によらず，バッテリ残量 0 から満充電までに 2700 ステップかかる．よって，エージェントは最大 3150 ステップの周期で巡回と充電を繰り返す．

清掃効率の評価指標である D_{t_s,t_e} は 3600 ステップごとに記録した．

4.2 清掃効率

図 1 の 2 つの環境で先行手法と提案手法での清掃効率 D_{t_s,t_e} を比較した．環境 (a)，(b) での結果をそれぞれ図 2，図 3 に示す． 両方の環境でどの時間帯においても既存と提案の手法の間に清掃効率の差がなく，改善も悪化もみられなかった．提案手法で報酬式を変更したことによって清掃効率が改善されないことが分かった．

4.3 エージェントの戦略

次に各目標ノード選択戦略を選んだエージェント数を調べた．1000 ステップごとに各エージェントの $Q^i(s)$ が最大である戦略を取得し，各戦略に対し最大の $Q^i(s)$ をもつ

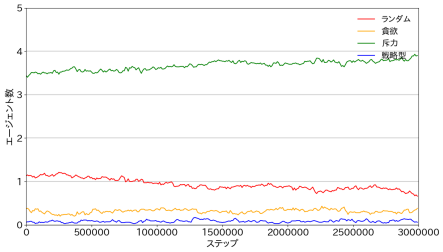

図 4 環境 (a)-既存手法における速さ 1 のエージェントの戦略ごとの台数推移

Fig. 4 Transition of the number of agents moving at a speed of 1 per strategy with existing method in Env. (a).

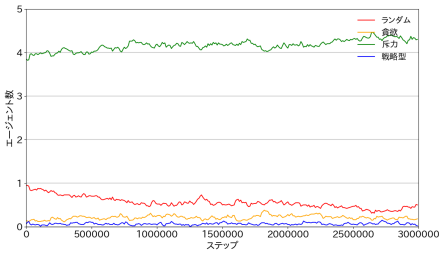

図 5 環境 (a)-既存手法における速さ 2 のエージェントの戦略ごとの台数推移

Fig. 5 Transition of the number of agents moving at a speed of 2 per strategy with existing method in Env. (a).

エージェント数を数えた．10000 ステップ，すなわちデータ 10 セットにつき 1 回，戦略ごとにエージェント数の平均をとり，さらにそれらの 10 試行での平均を計算してプロットした．我々はエージェントの速さによる戦略の違いを調べるため，エージェントの速さごとにグラフを作成した．本実験では表 1 の通り，ε-greedy 法の ε を 0.05 としている，そのため，エージェントは $Q^i(s)$ が最大となる戦略 s を 95%以上の確率で選択する．よって，本節で示す各図はエージェントが選択する戦略の傾向を表している．以下で環境ごとに結果を示し，考察する．

環境 (a) について，データを取得した各時刻に，各目標ノード選択戦略 s に対し最大の $Q^i(s)$ をもっていたエージェント数の推移を，既存手法の速さ 1，2 についてはそれぞれ図 4，5，提案手法の速さ 1，2 についてはそれぞれ図 6，7 に示す． 既存手法，提案手法ともに，斥力法が最大の $Q^i(s)$ となるエージェントが最も多く，速さ 1 のエージェントでは全 5 台中約 3.5 台，速さ 2 のエージェントでは約 4 台であった．時間の経過によって斥力法のエージェント数が 0.4 台程度増加し，その分ランダム法のエージェント数が減少した．また，速さ 1 のエージェントに比べて速さ 2 のエージェントの方が斥力法をおよそ 0.5 台多く選び，その分ランダム法は選ばない傾向があった．しかし，既存手法と提案手法の間に明らかな違いはみられなかった．環境 (a) はごみの発生確率がどの位置でも同じであり，ごみが多いノードを選ぶ貪欲法や近くにごみが多い場合に近くのノードを選ぶ戦略型目標決定法よりも，エージェントが分散して清掃する斥力法が選ばれやすいと考えられる．既存手法でも大半のエージェントが斥力法を選択していることから，斥力法の行動価値 $Q^i(s)$ が他の戦略の行動価値より差をつけて大きく，提案手法で報酬式を変えても，エージェントが選択する戦略の傾向が変化しなかったと考えられる．

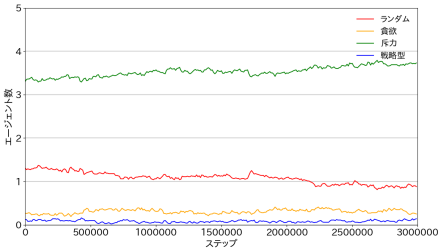

図 6 環境 (a)-提案手法における速さ 1 のエージェントの戦略ごとの台数推移

Fig. 6 Transition of the number of agents moving at a speed of 1 per strategy with proposed method in Env. (a).

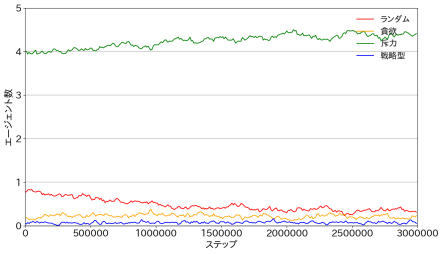

図 7 環境 (a)-提案手法における速さ 2 のエージェントの戦略ごとの台数推移

Fig. 7 Transition of the number of agents moving at a speed of 2 per strategy with proposed method in Env. (a).

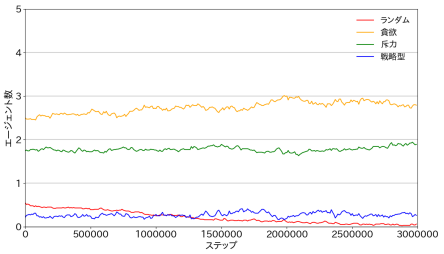

図 8 環境 (b)-既存手法における速さ 1 のエージェントの戦略ごとの台数推移

Fig. 8 Transition of the number of agents moving at a speed of 1 per strategy with existing method in Env. (b).

環境 (b) について，データを取得した各時刻に，各目標ノード選択戦略 s に対し最大の $Q^i(s)$ をもっていたエージェント数の推移を，既存手法の速さ 1，2 についてはそれぞれ図 8，9，提案手法の速さ 1，2 についてはそれぞれ図 10，11 に示す． 既存手法，提案手法両方において，速さ 1 のエージェントでは貪欲法が，速さ 2 のエージェント

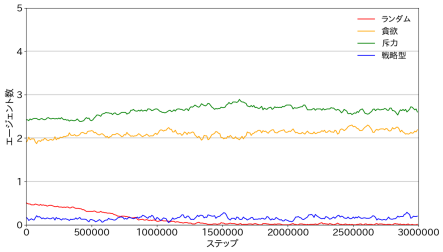

図 9 環境 (b)-既存手法における速さ 2 のエージェントの戦略ごとの台数推移

Fig. 9 Transition of the number of agents moving at a speed of 2 per strategy with existing method in Env. (b).

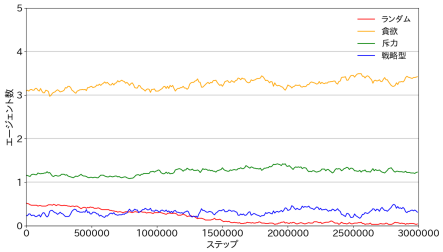

図 10 環境 (b)-提案手法における速さ 1 のエージェントの戦略ごとの台数推移

Fig. 10 Transition of the number of agents moving at a speed of 1 per strategy with proposed method in Env. (b).

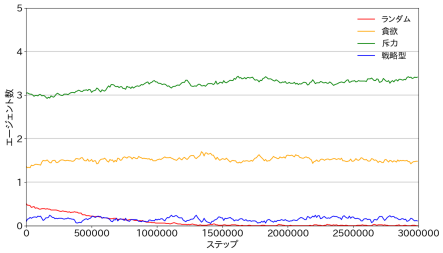

図 11 環境 (b)-提案手法における速さ 2 のエージェントの戦略ごとの台数推移

Fig. 11 Transition of the number of agents moving at a speed of 2 per strategy with proposed method in Env. (b).

では斥力法が最も多くなった．2 つの戦略間のエージェント数の差は，提案手法の方が既存手法よりも大きかった．既存手法，提案手法ともに環境 (b) においては選択した戦略の点では移動速度による分業ができていると言うことができる．環境 (b) はグリッドの外側にも内側にもごみがたまりやすい領域が存在し，内側の一部の領域はその中でも特にごみがたまりやすい．速さ 1 のエージェントは満充電の状態から移動できる距離が速さ 2 のエージェントの半分であり，充電基地から遠いノードを長時間回ることができない．よって，速さ 1 のエージェントは貪欲法に従って外側の領域よりも内側の最もごみが集まりやすい領域を目標ノードとして巡回し，速さ 2 のエージェントは内側の領域にエージェントが集中するため，斥力法によって外側の領

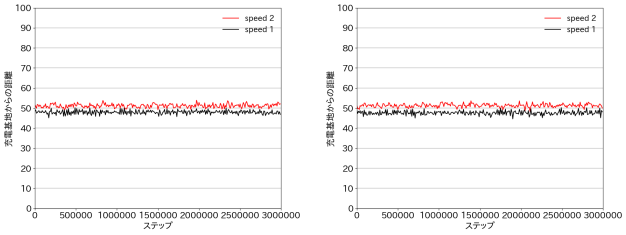

図 12 環境 (a) における速さごとの充電基地からの距離の推移．左が既存手法，右が提案手法のものである．

Fig. 12 Transition of average distance from charge base per agent's speed in Env. (a).

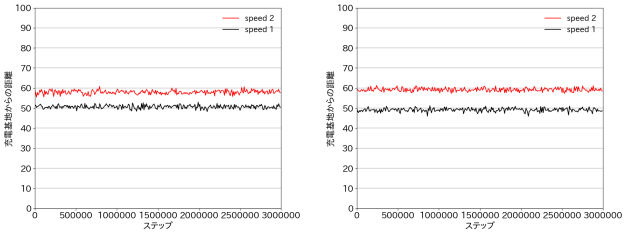

図 13 環境 (b) における速さごとの充電基地からの距離の推移．左が既存手法，右が提案手法のものである．

Fig. 13 Transition of average distance from charge base per agent's speed in Env. (b).

域などを巡回したと考えられる．

以上をまとめると，速さの違いによって戦略ごとに最大の行動価値をもつエージェント数が異なり，既存手法，提案手法の両方で戦略が異なるという点では分業が実現できたと言える．しかし，提案手法が既存手法と異なる結果を示したのは環境 (b) のみであった．大半のエージェントが特定の一つの戦略を選択するような環境では提案手法の報酬式で選択する戦略の傾向を変えることが難しいと分かった．報酬式 (10) の中で，速さ，充電基地からの距離による報酬の変動幅を大きくすることで環境 (a) でも選択する戦略の傾向を変えられる可能性があると考えられる．また，外側にも内側にもごみがたまりやすい領域をもつ環境 (b) において遅いエージェントが貪欲法，速いエージェントが斥力法を多く選択したことから，速い，遅いエージェントにそれぞれ斥力法，貪欲法を選択するように促すことで，充電基地からの距離による分業を実現できる可能性がある．

4.4 充電基地からの距離

速さごとにエージェントの充電基地からの距離の平均を 100 ステップごとに取得し，それらの 10000 ステップでの平均をとって 10 試行での平均値をプロットして推移を確認した．その推移を，環境 (a)，(b) についてそれぞれ図 12，図 13 に示す． 両方の環境で，速さ 2 のエージェントが速さ 1 のエージェントよりも充電基地からの距離の平均が 4 から 10 程度大きかった，これは速いエージェントは充電基地から速く遠ざかることができ，同じ稼働時間内で

より長い時間充電基地から離れた場所で活動したからである．環境 (b) では提案手法の方が既存手法より速さ 1，2 での距離の差が 3 程度大きかった一方，環境 (a) では提案手法と既存手法の間に違いがみられなかった．これは 4.3 で説明したように，環境 (a) においては提案手法と既存手法の戦略に違いがなかったためであると考えられる．一方，環境 (b) で提案手法での速さによる距離の差が大きくなったのは，既存手法と比べて速さ 1 のエージェントがより貪欲法を，速さ 2 のエージェントがより斥力法を選択するようになったからだと考えられる．環境 (b) においては，提案手法が戦略のみではなく充電基地からの距離の点でも分業を促すことができていると言える．

5. まとめ

本研究では複数エージェントによる巡回問題において，不均一な移動速度の複数エージェントが自らの移動速度に応じて分業することを目的に，既存手法を拡張して移動速度と距離によって戦略の報酬値を変化させる手法を提案した．エージェントの移動速度が 2 種類存在する状況で既存手法，提案手法それぞれで実験し，清掃効率，エージェントが選択する戦略，充電基地からの距離を比較した．その結果，エージェントの移動速度によって選択する戦略の傾向および，巡回ノードの充電基地からの距離に違いがみられた．一部の環境では，提案した手法は既存手法と比べて清掃効率，戦略の傾向，巡回ノードの充電基地からの距離に明らかな差がみられなかったが，環境の外側，内側の両方に汚れやすい領域をもつ環境では，戦略の傾向，巡回ノードの充電基地からの距離の点で提案手法が分業を促していることが分かった．

今後の課題として，提案手法がもたらす差異や分業が小さく，かつもたらす環境が限定されていることがあげられる．提案手法を用いたことで既存手法と比べて変化があった環境は本研究で使用した環境では一つだけであり，他の構造においても提案手法が効果をもたらす環境は少ないと考えられる．また，別のエージェントの速さやその構成で提案手法が効果をもたらすかどうかは不明である．多様な条件で移動速度に適した分業を促すように手法を改善する必要がある．既存手法の学習型目標決定法，その拡張である提案手法では別の戦略を選択肢に入れることができる．充電基地から近くまたは遠くのノードを我々は目標ノードとする戦略を新たに導入することで，本研究の提案手法で目的としていた充電基地からの距離による分業を実現できるのではないかと考えている．

参考文献

[1] M. Ahmadi and P. Stone. Continuous area sweeping: a task definition and initial approach. *ICAR '05. Proc., 12th Int. Conf. on Advanced Robotics, 2005.*, pp. 316–323, July 2005.

[2] M. Ahmadi and P. Stone. A multi-robot system for continuous area sweeping tasks. *Proc. 2006 IEEE Int. Conf. on Robotics and Automation, 2006. ICRA 2006.*, pp. 1724–1729, May 2006.

[3] D. Kurabayashi, J. Ota, T. Arai, and E. Yoshida. Cooperative sweeping by multiple mobile robots. *Proc. of IEEE Int. Conf. on Robotics and Automation*, Vol. 2, pp. 1744–1749, April 1996.

[4] K. Yoneda, C. Kato, and T. Sugawara. Autonomous learning of target decision strategies without communications for continuous coordinated cleaning tasks. *Proc. 2013 IEEE/WIC/ACM Int. Joint Conf. on Web Intelligence and Intelligent Agent Technologies*, Vol. 2, No. 4, pp. 216–223, 2013.

[5] K. Cheng and P. Dasgupta. Dynamic area coverage using faulty multi-agent swarms. pp. 17–23, 2007.

[6] J.J. Acevedo, C.B. Arrue, I. Maza, and A. Ollero. Distributed approach for coverage and patrolling missions with a team of heterogeneous aerial robots under communication constraints. *Int. Journal of Advanced Robotic Systems*, Vol. 10, No. 1, p. 28, 2013.

[7] P. Dinnissen, S. N. Givigi, and H. M. Schwartz. Map merging of multi-robot slam using reinforcement learning. pp. 53–60, Oct 2012.

[8] 杉山歩未, 菅原俊治. 動的環境におけるマルチエージェント巡回清掃の自律的な戦略学習の提案. 電子情報通信学会論文誌, Vol. J98-D, No. 6, pp. 862–872, 2015.

[9] D. Hennes, D. Claes, W. Meeussen, and K. Tuyls. Multi-robot collision avoidance with localization uncertainty. *Proc. of the 11th Int. Conf. on Autonomous Agents and Multiagent Systems*, Vol. 1, pp. 147–154, 2012.