動的に変化する環境の中で自己位置を推定する自律走行ロボット

西村 真衣

移動ロボットの自律移動システムは搭載センサによる環境認識と地図構築、その地図上での自己位置認識に基づく経路計画によって構成される。静的な環境においては事前に地図構築を行い、環境センサ情報を登録地図と照合することにより比較的安定して移動システムを動作させることができるが、人が介在するなど動的な環境においては地図構築の観点からはノイズとなる移動体を如何に処理するかが課題となる。本稿では自律移動システムのうち特に動的環境を対象とした自己位置推定に焦点をあて、移動体が存在する環境下における主なアプローチを俯瞰すると共に、我々が提案する移動体のみを利用する新たな移動軌跡復元手法の枠組みについて述べる。

Mobile Robot Navigation in Densely Crowded Environment

NISHIMURA Mai

The autonomous mobile robot (AMR) system consists of environment mapping using equipped sensors, self-localization and path planning on the top of the map. In static environment, where the layout of objects is fixed and no dynamic objects appear in the map, the system can build the whole environment map in advance. However, when it comes to dynamic environments, the AMR system is required to build the map sequentially by handling dynamic obstacles in real-time. In this document, we especially focus on the self-localization system in the dynamic environment. We first briefly review fundamentals of localization system based on the multi-view geometry and introduce its extensions to incorporate dynamic points and obstacles. Moreover, towards navigating in highly congested scenarios, we propose a novel self-localization framework that depend only on dynamic objects in the observed images.

1. まえがき

移動ロボット（モバイルロボット）、またその土台となる自律移動システムは工場内や公共施設における運搬、警備、物流、清掃など多様なシーン、用途で活用されている。一方で工場内のように経路、障害物を固定できる環境と比較して、人が介在するなど特に動的に大きく変化する環境へのモバイルロボットの導入は未だ課題が多い。自律移動システムは搭載センサによる周辺環境の認識と地図構築、構築した地図上での自己位置認識と経路計画によって構成されるが、その全てのプロセスにおいて移動体の存在を考慮する必要がある。第一に地図構築において、静的な環境では事前に地図構築を行うことが可能であるが、移動体を含む環境地図は刻一刻と変化するため、固定の静的な背景と移動体による動的な前景を分離可能にした環境地図を逐次更新しながら構築し、走行可能領域を算出する必要がある。更に、その構築した地図上で自律走行を行うには、変化する環境地図上での自己位置及び障害物となる移動体の認識が不可欠となる。本稿では、このような動的環境下を対象とした自律移動システムの中で特に自己位置推定に着目し、既存のアプローチを俯瞰すると共に、移動体のみを利用した新たな自己軌跡復元の枠組みについて述べる。

移動ロボットでは、IMU（Inertial Measurement Unit）、LiDAR 等多様なセンサが利用されているが、本稿では特に単眼 RGB 画像センサで構成される自律移動システムについて解説する。RGB カメラを用いた地図構築及び自己位置推定は静的なランドマークを複数の視点から観測し、多視点幾何による幾何学的な拘束式を解くことによりカメラの位置パラメータを逐次算出することが基本となる。そのため、環境地図及び自己位置推定は静的なシーンの観測

Contact : *NISHIMURA Mai*　mai.nishimura@sinicx.com

が前提とされ、歩行者や自転車、自動車など移動体が含まれる環境ではそれらを外れ値として、又は意味レベルで障害物として認識し、除外する方法が取られてきた[1,2]。更に、移動体が多く存在するシーンでは安定した自己位置推定を実現するため、移動体であるオブジェクトも同時にランドマークとして複合的に最適化対象とするアプローチも提案されている[3,4]。しかしながら、これらはいずれも静的なランドマークの観測をベースとした従来の手法の拡張として提案されており、静的なランドマークが安定してトラッキングし続けられることが前提となる。

以上で述べた通り、従来の自律移動システムは周辺環境が静的で変化しないことが前提であり、動的な環境の変化に対応するための拡張手法群もシーン全体の中で静的なランドマークが支配的である環境を対象としている。では、雑踏など移動体の存在がシーンの内で支配的であり、遮蔽により静的なランドマークの安定した追跡が不可能な状況下でも尚自律走行ロボットを運用し続けるにはどのような方法が取れるだろうか。本稿の後半では更に、そのような従来のアプローチが機能しない“超”動的な環境における自己位置推定の新しい枠組みについて紹介する。

2. 多視点幾何の基礎

2.1 多視点撮影による幾何パラメータの推定

移動ロボットでは以下ⅰ）～ⅳ）のプロセスの反復によって移動しながら環境地図と自己位置を逐次更新していく。

ⅰ）2 視点でのランドマークの検出・マッチング
ⅱ）ランドマークの 3 次元位置及びカメラ位置姿勢の復元
ⅲ）復元したランドマークを環境地図へ登録
ⅳ）新たな観測点において再度ランドマークを検出、登録されたランドマークとの対応関係から、カメラ位置姿勢を推定
ⅴ）新規に観測されたランドマークの 3 次元位置を復元し、環境地図へ登録

このように地図構築（Mapping）と自己位置推定（Localization）は相互に依存しているため、自律走行におけるロボットの自己位置推定技術は Simultaneous Localization and Mapping（SLAM）と呼称される。従来の SLAM システムでは、観測する環境が静的で変化しないランドマーク（特徴点）が存在することを前提とし、複数フレームで共通のランドマークが追跡可能である必要がある。

図 1 に多視点幾何の概念図を示す。3 次元空間中のランドマーク P は各視点（View1,View2）において画像平面上の点 p, p' として投影される。この画像平面上における対応点を用いて、視点間のカメラ位置姿勢を回転行列 R, 並進移動ベクトル t を算出する。

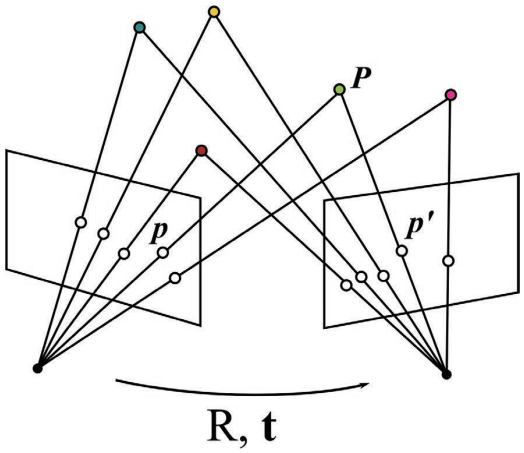

図 1　共通のランドマークの観測によるカメラ幾何パラメータ推定

2.2 バンドル調整

バンドル調整[17]とは、図 1 におけるランドマーク P の 3 次元位置とカメラを結ぶ光線束（Bundle）を対応する撮影画像上の特徴点 p の投影関係を利用して調整し、その背後にある幾何パラメータを推定・最適化するための方法である。実用上、計算過程において画像平面上における特徴点 p の検出座標には観測誤差が含まれるため、ランドマーク P の 3 次元位置及びカメラ位置姿勢パラメータの推定結果はその誤差の影響を受けたものとなる。バンドル調整ではこのような観測誤差の混入を前提とし、推定された各カメラ姿勢 R_i, t_i について元の観測された点 p_i と 3 次元点 P_i の投影関数 $f(P_i \mid R_i, t_i)$ による再投影誤差

$$\min_{R_i, t_i} \sum \left| p_i - f(P_i \mid R_i, t_i) \right|_2^2 \qquad (1)$$

を最小二乗法によって最小化することで最適なパラメータを求める。なお、最小二乗法は誤差に対し正規分布のノイズを仮定した場合の最尤推定に相当する。多視点幾何を用いたカメラの位置姿勢推定は観測を増やすごとに各処理の誤差が蓄積していくため、過去に推定された位置姿勢パラメータ群を局所的なグループ毎（local Bundle Adjustment, local BA）またはループが検出されたタイミングで過去に推定された全ての位置姿勢に対して最適化を行う（global Bundle Adjustment, global BA）。

3. 動的な環境下における自己位置推定

3.1 ロバスト推定手法

前章までで述べたように従来の自己位置推定技術は静的なランドマークを多視点で観測することを基本として構成されている。従って、観測された特徴点の一部に動的な点

群が含まれた場合、それらをまず外れ値として処理するアプローチが取られてきた。その最も代表的な手法がロバスト推定手法である。ロバスト推定とは与えられた観測値、予測値に外れ値が含まれていることを前提とし、その誤差影響を抑えることを目的とした手法群である。代表的な手法として Random Sample Consensus（RANSAC）や Least Median of Squares（LMedS）、M-estimator 等がある。2.2 節で触れた最小二乗法では、目的値 y に対する予測値 $f(x)$ の二乗誤差最小化を行うが、観測に含まれる誤差の分布が正規分布に従うことを仮定していた。

$$LMS = \min_{\Delta x} \sum \epsilon^2, \qquad \epsilon = |y - f(x)| \tag{2}$$

これに対し、M-estimator は誤差基準 ϵ に対して誤差重み関数 ρ を設定し、重み付き誤差を最小化するアプローチである。

$$M = \min_{\Delta x} \rho(\epsilon) \tag{3}$$

重み関数としては外れ値に対してより小さな重みを与えるような偶関数が選ばれることが多く、特徴点ベースで代表的な Visual SLAM 手法である ORB-SLAM[5,6] では Huber のコスト関数を使用している。また、一般にはデータに対するノイズの分布は未知であることから、学習データに対して最適なこのロバスト推定手法における誤差関数 ρ を学習させる手法[7] も近年提案されている。このようなロバスト推定に基づくアプローチは主に誤検出や僅かな移動体によって生じる外れ値を対象としており、移動体が多く存在するようなシナリオには適用できない。

3.2 Dynamic SLAM

近年の移動体を含む動的なシーンを対象とした SLAM システムの研究群（Dynamic SLAM）では、大きく二つの方向性において従来の静的シーンを前提とした SLAM システムを拡張している。まず 1 つは、環境マップを離散的な点群として扱うのではなく、Semantics を考慮して Object レベルでの処理を行う Object-aware なシステムを構築している点である。最も単純な実装として、Semantic レベルで検出した移動オブジェクトを個別にトラッキングしながらフィルタリング処理によって環境地図を構成するランドマークから取り除く方法が提案されている[1,2]。更にもう 1 つは、静的なランドマークと同時に移動体であるオブジェクトとその動きを外れ値として除外せず、最適化対象として含めるという点である。

移動体を同時に最適化対象に含める Dynamic SLAM では、静的なランドマークの検出と同時にオブジェクトレベルでの検出・追跡処理が実行される。つまり、静的シーンを扱う SLAM システムが観測フレーム間でマッチング可能な静的な特徴点のみを対象とするのに対し、Dynamic SLAM では特徴点及びオブジェクトのフレーム間対応付けを同時に行い、因子グラフとしてバンドル調整のパイプラインに組み込むアプローチをとっている[3,4]。

4. 移動体のみを利用する自己位置推定

既存の静的なランドマーク観測に基づく SLAM の拡張として記述される Dynamic SLAM では、移動体を含むシーン全体の復元を実現している一方で、移動体と同時に常に安定して静的なランドマークを追跡し続ける必要があった。従って歩行者で混雑した路上など移動体が非常に多く含まれ、相互遮蔽により静的なランドマークの安定した追跡が不可能な状況下において自己位置推定を行うことは極めて難しい。では見方を変えて、静的なランドマークの観測に依存せず、移動体のみを用いてカメラの自己運動を含めた周辺環境を復元することはできないだろうか。つまり、図 2 に示すように一人称視点での周辺歩行者の動きの 2D 観測から、俯瞰視点でのカメラ自己運動及び周辺歩行者の移動軌跡を復元する問題を考える。

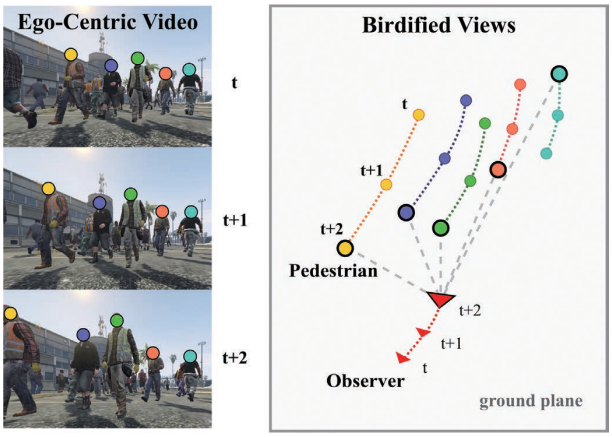

図 2　一人称観測に基づく自己及び周辺歩行者の軌跡復元

各タイムステップ t において、歩行者 1, ..., K の一人称映像に基づく相対的な観測 $Z^t_{1:K} = \{z^t_1, z^t_2, ..., z^t_K\}$ が与えられたとき、その俯瞰視点上での位置 $X^t_{1:K} = \{x^t_1, x^t_2, ..., x^t_K\}$ 及び観測カメラの位置 x^t_0 を同時復元するとする。移動体が大きさ（身長）

$$\begin{bmatrix} \tilde{x}_k \\ \tilde{y}_k \\ \tilde{z}_k \end{bmatrix} = \frac{f h_k}{l_k} R_x\left(-\frac{\pi}{2}\right) A^{-1} \begin{bmatrix} u_k \\ v_k \\ 1 \end{bmatrix} \tag{4}$$

が限られる人物であると仮定すると、画像平面上における人物検出位置 $[u_k, v_k]^\top$ に対し、観測カメラに対する相対的な 3D 位置 $[\widetilde{x_k}, \widetilde{y_k}, \widetilde{z_k}]^\top$ は、カメラ内部行列 A, 焦点 f を用いて以下に示す逆投影で表現される。ここで、$R_x(\theta)$ は x 軸周りでの角度 θ の回転、h_k, l_k はそれぞれ人物の 3 次元空間での身長、画像平面における観測上の身長である。

実環境での人物の身長が平均 μ_h、分散 σ_h のガウス分布

に従う[9] とすると、観測モデルは逆投影線上のガウス分布として記述される。

$$z_k^t \sim p(z_k^t \mid x_k^t; h_k) = N(\mu_h, \sigma_h^2) \tag{5}$$

また、観測される歩行者が群として共通の移動モデルに沿って移動していると仮定すると、カメラ及び歩行者の移動軌跡 $X_{0:K}^t$ は観測モデルである尤度分布 $p(z_k \mid x_k)$ 及び移動モデルについての事前分布 $p(x_k)$ を用い、以下の事後分布を最大化させることによって求められる。

$$p(X_{0:K}^t \mid Z_{1:K}^t, X_{0:K}^{t-1}) \propto p(X_{0:K}^t \mid X_{0:K}^{t-1}) p(Z_{1:K}^t \mid X_{0:K}^t, X_{0:K}^{t-1}) \tag{6}$$

ここで、周辺歩行者の推定位置 $X_{1:K}^t$ を固定すると、観測カメラの移動量 $\Delta x_0^t = [\Delta x_0^t, \Delta y_0^t, \Delta\theta]$ は以下の MAP 推定により求められる。

$$\Delta \hat{\boldsymbol{x}}_0^t = \underset{\Delta x_0^t \in \mathbb{R}^3}{\arg\max}\, p(\boldsymbol{x}_0^t \mid \chi_0^{t-\tau:t-1}) \prod_k p(\boldsymbol{x}_k^t \mid \hat{\chi}_k^{t-\tau:t-1}, \Delta \boldsymbol{x}_0^t) p(\boldsymbol{z}_k^t \mid \boldsymbol{x}_k^t, \Delta \boldsymbol{x}_0^t), \tag{7}$$

ここで、$p(x_0^t \mid X_0^{t-\tau:t-1}), p(x_k^t \mid X_k^{t-\tau:t-1})$ は τ フレーム分の観測におけるそれぞれカメラ、周辺歩行者の移動モデルである。周辺歩行者の移動軌跡は上記で推定されたカメラ位置 Δx_0^t を固定することにより、同様に以下式により求められる。

$$\hat{\chi}_{1:K}^t = \underset{x_k^t \in \chi_{1:K}^t}{\arg\max} \prod_k p(\boldsymbol{x}_k^t \mid \chi_k^{t-\tau:t-1}, \Delta \hat{\boldsymbol{x}}_0^t) p(\boldsymbol{z}_k^t \mid \boldsymbol{x}_k^t, \Delta \boldsymbol{x}_0^t) \tag{8}$$

この観測カメラ移動量の推定と歩行者位置推定は相互に依存しているため、交互に推定量を固定しパラメータの逐次更新を繰り返すことにより、カメラ及び歩行者軌跡の復元を行う。

5. データセット

動的環境での自己位置推定の評価を行うデータセットは 3 次元シーンをキャプチャしたデータに 2 次元の動きを含む移動体を撮影画像に 2D 合成したデータセット[8] 等が提案されているが、歩行者により混雑した環境を対象とし、その移動軌跡とカメラ自己運動の復元を目的としたデータセットは存在しない。また、リアルな歩行者の軌跡データは俯瞰視点映像で撮影されたものが殆どであり[10-12]、一人称視点での群衆の動きを捉えたデータセットは希少である。そこで、今回は人工的に生成又は群衆の俯瞰視点映像から抽出した軌跡データを仮想カメラモデルによって一人称映像に投影し、疑似的に生成した一人称視点の観測と俯瞰視点での軌跡データをペアとしたデータを作成した。

5.1 シミュレーションデータセット

仮定する移動モデルが既知のケースにおける提案法の性能を検証するため、Social Force Model [13] によって人工的に生成した軌跡データを 10 ～ 50 人の群衆内に含まれる人数毎に作成した。仮想カメラモデルの内部パラメータは既知とし、歩行者の身長はガウス分布 $h_k = N(\mu_h, \sigma_h^2)$, $\mu_h = 1.70[m]$, $\sigma_h \in [0.00, 0.07]\,[m]$ に従ってサンプリングしたデータを生成した。

5.2 リアルな歩行者軌跡を利用したデータセット

Motion Model が未知の軌跡データに対する提案法の有効性を確認するため、公開の群衆データセットである ETH[10]、UCY [11] から映像中に含まれる群衆人数に応じて軌跡データを抽出し、歩行者の 1 人に仮想カメラをマウント・その視点における疑似的な投影データを俯瞰視点軌跡とペアとしたデータを作成した。

5.3 写実的なゲーム映像を利用したデータセット

5.1、5.2 節で作成したデータはいずれも仮想的なカメラモデルを用いた疑似投影であり、歩行者間の相互遮蔽は考慮されていない。そこで、コンピュータビジョン分野での CG データセットとして広く利用されている Grand Theft Auto V（GTAV）[14] を用い、人物軌跡データとペアとなる写実的な一人称映像を生成した。GTAV では Script Hook V[15] というライブラリを用いて GTAV 内のネイティブ関数を呼び出す任意のプラグインを記述することができ、指定のシーンにおいて人物配置や深度画像を逐次取得しながら、各エージェントについてプログラムされた軌跡通りの行動を実行することが可能である。図 3 に GTAV によって作成した軌跡データのサンプルを示す。

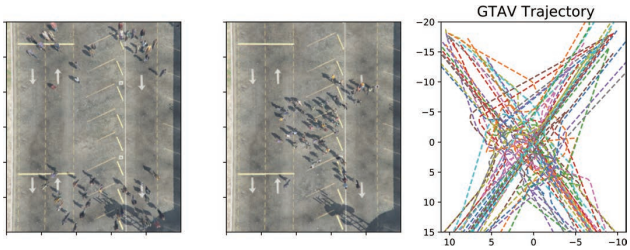

図 3　GTAV で作成した軌跡データサンプル

6. 実験

6.1 シミュレーションデータでの評価

5.1 節で作成した既知の移動モデルを利用して構築したデータセットにおいて、提案手法の評価を行った。図 4 にシミュレーション結果を示す。$\Delta r, \Delta t$ はそれぞれ自己運動による時刻 $t-1 \sim t$ での回転角及び移動量、$\Delta\tilde{x}, \Delta x$ は周辺歩行者のカメラに対する相対位置、観測カメラの運動を含めた絶対位置の推定結果である。歩行者の身長の分散 σ_h が大きいほど推定位置のエラー率は上昇するが、群衆の人数が増加するほどカメラの自己運動の推定精度は高くなり、その観測カメラ位置に基づく周辺歩行者の位置推定精度も高くなる傾向が認められた。

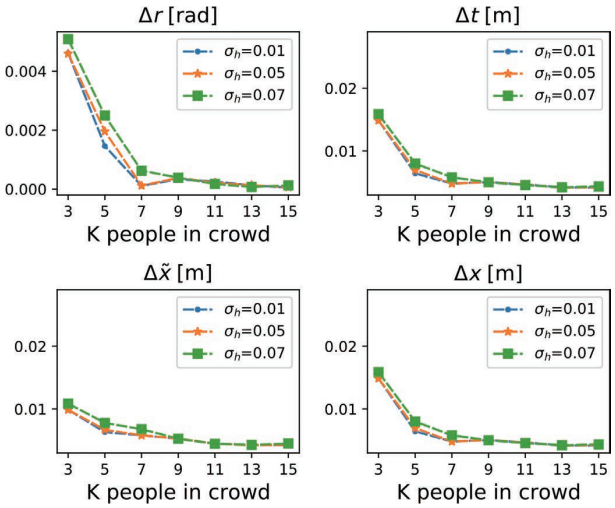

図 4　シミュレーションデータにおける提案法の推定結果

6.2 リアルな移動軌跡データセットでの評価

5.2 節で作成した未知の移動モデルによる群衆の軌跡データセットを利用して構築したデータセットにおいて、提案手法の評価を行った。図 5 に提案法によって推定されたカメラ及び歩行者位置の確率分布の可視化結果を示す。カメラ軌跡は赤、歩行者軌跡はグレーで描画され、上段（GT）が正解データ、下段（estimated）は推定分布に相当する。数値実験結果は割愛するが、$t \sim t+3$ のように歩行者数が少人数の際は推定される位置分布の裾が拡がるため曖昧性が高くなり、$t+4 \sim t+9$ のように歩行者人数が一定以上のケースでは分布の尖度が向上し、観測する移動体が多くなるに従って位置推定は安定して機能する特性が認められた。

6.3 GTAV により生成した一人称映像データによる評価

GTAV により生成した一人称映像に対し、MOT-16 によって事前学習された既存の Multi-Object Tracker（MOT）[16] を用いて画像平面上での人物位置に相当する Bounding Box を抽出し、そのトラッキング結果に基づいて俯瞰視点上でのカメラ・歩行者移動軌跡復元を行った。図 6 に提案法の推定結果を示す。より実用的な場面を意識し、MOT の出力を軌跡復元の入力として用いた場合においても、混雑した環境下において提案法は安定したトラッキング結果を示した。

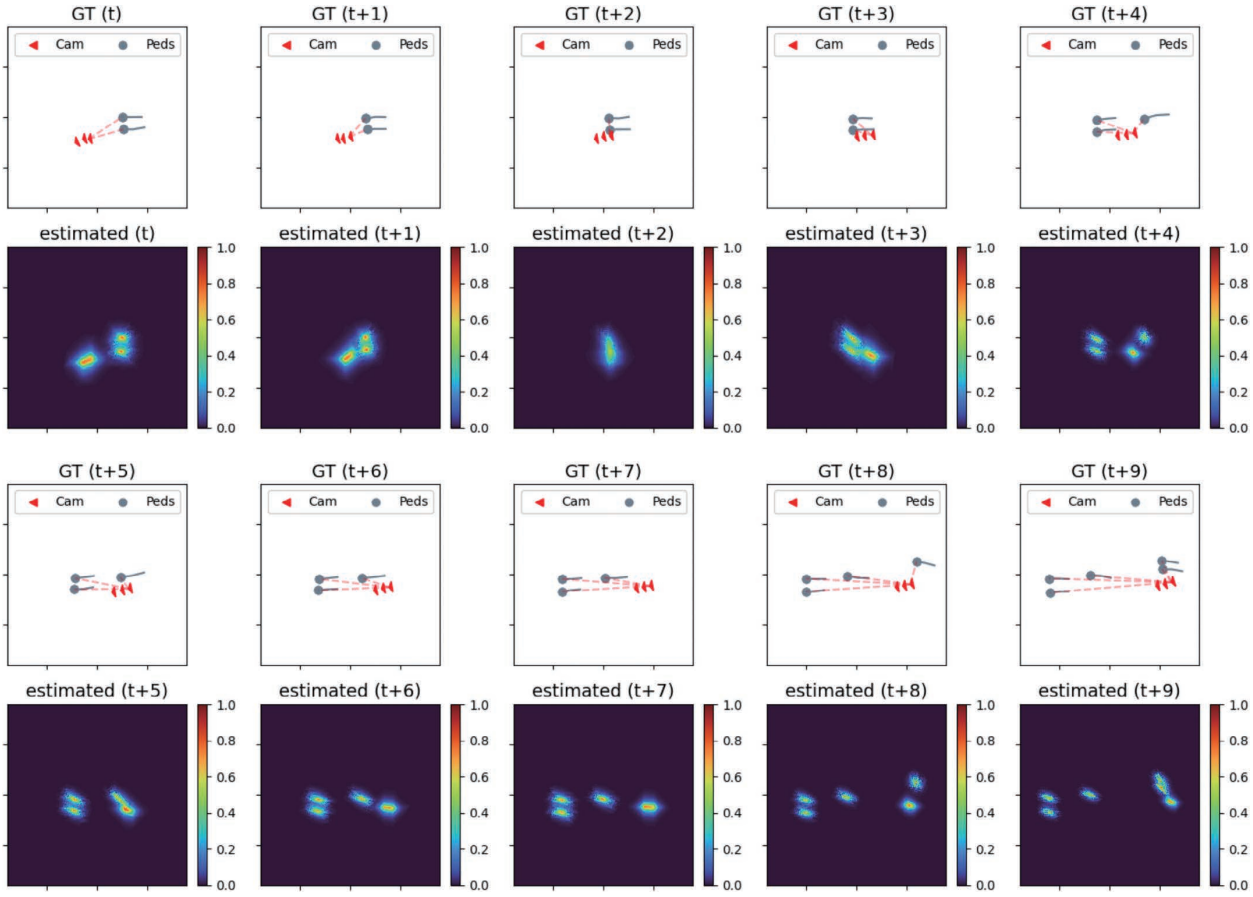

図 5　リアルな軌跡データに対するカメラ及び歩行者位置の確率分布推定結果

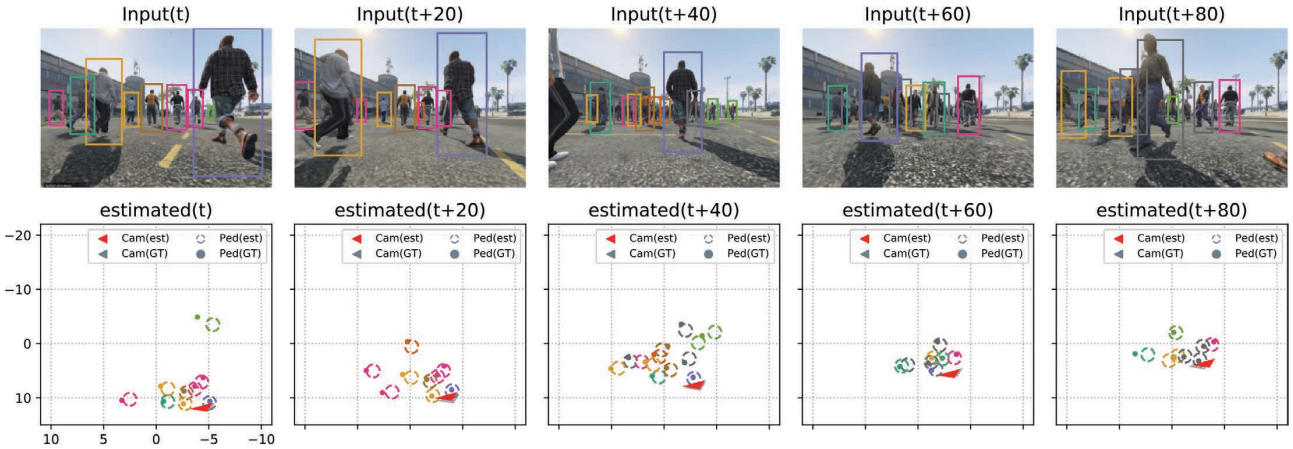

図 6　CG により生成した一人称映像（上段）に対する俯瞰視点移動軌跡推定結果（下段）

7. むすび

本稿では、多視点幾何に基づく移動ロボットの自己位置推定の基礎を俯瞰するとともに、動的な環境を対象とした拡張手法について紹介した。また後半では更に、静的なランドマークの観測を基本とした従来手法では扱うことができなかった混雑環境下において、移動体のみをランドマークとして用いることにより移動ロボット及び歩行者軌跡を復元する取り組みを提案した。移動体のみを利用するアプローチはテクスチャが乏しい背景環境においても適用することが可能であり、これまで従来法の枠組みで扱うことが不可能であった様々なシーンへ応用を広げることが期待される。

参考文献

1）Bescos, B. et al. DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes. IEEE Robotics and Automation Letters. 2018, Vol. 3, No.4, p.4076−4083.

2）Yu, C. et al. “DS-SLAM: A semantic visual SLAM towards dynamic environments”. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems（IROS）. 2018, p.1168−1174.

3）Henein, M. et al. “Dynamic SLAM: The need for speed”. 2020 IEEE International Conference on Robotics and Automation（ICRA）. 2020, p.2123−2129.

4）Huang, J. et al. “Clustervo: Clustering moving instances and estimating visual odometry for self and surroundings”. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, p.2168−2177.

5）Mur-Artal, R.; Jose, M. M. M.; Tardos, J. D. ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Transactions on Robotics. 2015, Vol.31, No.5, p.1147−1163.

6）Mur-Artal, R.; Tardós, J. D. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE Transactions on Robotics. 2017, Vol.33, No.5, p.1255−1262.

7）Lv, Z. et al. “Taking a deeper look at the inverse compositional algorithm”. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, p.4581−4590.

8）Lv, Z. et al. “Learning rigidity in dynamic scenes with a moving camera for 3d motion field estimation”. Proceedings of the European Conference on Computer Vision（ECCV）. 2018, p.468−484.

9）Luo, Y. et al. “Where, What, Whether: Multi-modal learning meets pedestrian detection”. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, p.14065−14073.

10）Pellegrini, S. et al. “You’ll never walk alone: Modeling social behavior for multi-target tracking”. 2009 IEEE 12th International Conference on Computer Vision. 2009, p.261−268.

11）Lerner, A.; Chrysanthou, Y.; Lischinski, D. Crowds by example. Computer Graphics Forum. 2007, Vol.26, No.3, p.655−664.

12）Wen, L. et al. “Detection, Tracking, and Counting Meets Drones in Crowds: A Benchmark”. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, p.7812−7821.

13）Mehran, R.; Oyama, A.; Shah, M. Abnormal crowd behavior detection using social force model. 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009, p.935−942.

14）Rockstar Games. “Rockstar Games”. https://www.rockstargames.com,（参照 2021−09−17）.

15）Script Hook V. “Script Hook V”. http://www.dev-c.com/gtav/,（参照 2021−09−17）.

16）Wang, Z. et al. “Towards real-time multi-object tracking”. Computer Vision−ECCV 2020: 16th European Conference. Proceedings, Part XI 16. Springer International Publishing, 2020, p.107−122.

17）岡谷貴之．バンドルアドジャストメント．情報処理学会研究報告．2009, Vol.2009-CVIM-167, No.37, p.1−16.

執筆者紹介

西村 真衣　NISHIMURA Mai
オムロン サイニックエックス株式会社
リサーチアドミニストレイティブディビジョン
専門：コンピュータビジョン、GPGPU
所属学会：IEEE

本文に掲載の商品の名称は、各社が商標としている場合があります。