市街地環境下における画像場所認識を用いた自己位置推定法

○三橋雅仁(明治大) 高橋佑弥(明治大) 黒田洋司(明治大)

Visual Localization System in Outdoor Environments using Appearance Based Place Recognition

*Masahito MITSUHASHI (Meiji Univ.), Yuya Takahashi (Meiji Univ.)
and Yoji KURODA (Meiji Univ.)

Abstract— In this paper, we propose a mobile robot localization system using appearance based place recognition. Our system uses appearance based place recognition and dead reckoning with Wheel Odometry. We use the theory of the observation likelihood estimation introduced from Fast Appearance Based Mapping (FAB-MAP) for appearance based place recognition. Images collected by a robot produce Bag-of-Words representation and Chow Liu trees. This representation is based on quantized SURF detectors/descriptors. A certain place can be recognized when the collected image matches a reference image coming from the same place in scene database. In addition, problems of appearance based place recognition techniques are solved by using sensor fusion. Experimental results are presented in an outdoor environment. The proposed localization system can achieve the average position error within 25m without GPS in an outdoor environment.

Key Words: Localization, Place Recognition, FAB-MAP, Divided Difference Filter

1. 緒言

近年、科学技術の発展に伴ない、屋内外における自律移動ロボットへの期待が高まり、研究が盛んに行われている。自律移動ロボットにとって、ロボット自身が場所の認識を行い、自己位置の推定を行う事は必要不可欠な技術の一つである。

市街地環境内において自己位置推定を行うには、画像情報を用いた場所認識が有効である事が知られている[7]。近年では、Bag-of-Wordsを用いた場所認識法が提案され、高速な場所認識が行えるようになった[6]。また、より高精度な場所認識を実現するために、visual words同士の繋がりも考慮したFast Appearance Based Mapping (FAB-MAP)が提案された[1]。この手法はこれまでの手法と比べ、より高精度な推定を可能にした。しかしながら、画像情報を用いた場所認識は、画像間の対応を得るだけのものであり、実際のロボットのナビゲーションに必要なロボットの位置や姿勢を推定するシステムはこれまであまり提案されてこなかった。

そこで本論文では、この場所認識とリファレンスデータを組み合わせて、高精度でロバスト性の高い自己位置推定を実現するシステムを提案する。リファレンス画像に位置座標を組み合わせて場所認識を行う事により、ロボットの位置と姿勢の情報を得る。

しかし、画像情報を用いた場所認識の大きな問題として、誤認識がある。誤認識の原因として、大きく三つがあげられる。一つは、草むらなど非常に類似した景色が多数存在する事である。その為、こうした画像では、同じvisual wordを多数観測してしまい、誤った認識を引き起こす。二つ目は、車や人などの移動体が画像に写りこむ場合である。このような場合、人や車などの移動体から、visual wordが観測されてしまい、画像間の類似度が低くなる。三つめは、取得画像に車や人のみが写りこむ場合である。この場合、別の場所

で同じ車や人が写りこんだ画像とマッチングをしてしまう。これらの理由より、その画像を取得した位置と全く違う場所にも関わらず、同じ場所であると誤認識してしまう事もある。

これらの問題から、画像情報による場所認識を自己位置推定に用いた場合、誤認識を取り除いたとしても、その間、場所認識からの観測値は得ることができない。つまり、画像情報による場所認識を自己位置推定に使用するには、このような誤認識や、それによって長時間観測値が得られない場合でも、安定した自己位置推定システムが必要である。

そこで本研究では、これらの画像情報による場所認識の問題を解決するための自己位置推定法を提案する。場所認識とデッドレコニングを組み合わせ、ロボットの動作モデルも考慮する事で場所認識における誤認識を軽減する。また、センサ統合をする事により、場所認識からの観測値が長時間得られない場合でも、安定した推定を可能とする。

本論文では、システムの有効性を示すために、つくばチャレンジ2010のコースで実験を行う。本手法による自己位置推定の結果とGround truthとを比較し、精度を検証する。

2. システム構成

本研究で提案するシステムでは、場所認識から得られる座標とデッドレコニングとを、センサ統合を行う事で、自己位置推定を行う。

本研究で提案する手法では、事前に経路を走行させ、データの取得を行う。この際に取得した画像は、画像データベースに登録してリファレンスデータとして用いる。また、本手法では単眼画像を用いるため、複雑な三次元復元やカメラの内部パラメータも必要としない。さらに、このリファレンスデータに、取得を行った際の位置情報を持たせる事で、場所認識により位置

Fig.1 An example of the place recognition using Google Street View. Top images are captured in the experiment. Bottom images are the images from Google Street View.

を特定できるようにする。この際の位置情報は、精度の良い GPS からの観測値を利用する。この理由として、一度だけ正確な位置情報を持つ画像を取得しておく事で、以降何度も使用できるためである。また、近年では、Google Street View[12] のように位置情報を持つ画像が容易に入手できる事も挙げられる。図 1 には、Google Street View をリファレンスデータに用いた場所認識の例を示す。図 1 で示すように、リファレンスデータを Google Street View で行った場合でも、本手法により認識が行える事を確認している。

3. 画像情報による場所認識

3-1 FAB-MAP

FAB-MAP は、Cummins らによって提案された手法で、画像情報のみを用いて、式 (1) のベイズ則によって画像間の類似度の計算を行う [1]。得られた類似度の中で最も高い類似度の画像を選択する事で、その画像を取得した場所と同じ場所であると認識が行える。

$$p(L_i|Z^k) = \frac{p(Z_k|L_i)p(L_i|Z^{k-1})}{p(Z_k|Z^{k-1})} \quad (1)$$

ここで、 L_i は位置、 Z_k はある時刻 k での画像情報の観測を表す。また、 Z^k は時刻 k までに取得した全観測値である。さらに、 $p(L_i|Z^{k-1})$ は、事前の情報 (Location prior) で、 $p(Z_k|L_i)$ は、観測値からの尤度 (Observation Likelihood) である。 $p(Z_k|Z^{k-1})$ で正規化を行う。本研究では、FAB-MAP の (Observation Likelihood) の考えを利用し、場所認識に利用する。詳細は文献 [1] を参照されたい。

この手法には、画像から SURF 特徴量を検出し、Bag-of-Words の考え方を用いて画像を表現する。また、Chow-Liu trees アルゴリズムを用いて、それぞれの word 同士の繋がりも加味して確率計算を行う。各手法について以降の章で説明を行う。

3-2 SURF 特徴量

画像情報からロバストに情報を得るには、特徴点の抽出が一般に知られている。本システムに用いる SURF は、特徴点の検出と特徴量の記述を行うアルゴリズムである。検出した特徴点に対して、画像の回転・スケール変化・照明変化等にロバストな特徴量をベクトルとして記述できる。同様の SIFT に比べ、計算時間も短いため、本システムでは SURF を用いる。また、自律移動ロボットにとっては、常に道路の上を走っているために、取得する画像が回転変化している事は考えられない。そのため、本システムでは画像の回転に不変な特性を取り除いた U-SURF を用いる [2]。また、U-SURF は通常の SURF と比べてより計算が早く行える。

3-3 Bag-of-Words

画像を取得し、大量の画像から類似度を計算する場合、大量の特徴量と比較しなければならず、計算時間が増大してしまう。こうした問題を解決するために本システムでは Bag-of-Words を導入する [5]。これは、特徴量ベクトルの量子化を行い、画像を visual word の集合として表現を行う方法である。

visual words の作成は、学習画像から特徴量を記述し、全学習画像の全特徴量を k-means 法によりクラスタリングを行う。そして、各特徴量を visual word として、vocabulary を作成する。こうして得られる各 word の代表ベクトルから、画像を word の観測として表現する。実際の処理においては、非常に似ている特徴同士が同じ word に属するように、クラスタリングを行う。

3-4 Chow Liu trees

取得した画像と以前の画像との尤度は、word 一つ一つを単体として近似する Naive Bays が一般的に用いられる。本システムでは、よりロバストな認識を行うために、word 同士の存在確率の繋がりを加味して近似を行う Chow Liu trees アルゴリズムを用いる [3]。式 (2) を用いる事で各 word の相互情報量の計算を行い、最も相互情報量の高い組み合わせを選択する事で、Chow Liu trees が得られる。

$$I(z_i, z_j) = \sum_{z_i \in \Omega, z_j \in \Omega} p(z_i, z_j) \log \frac{p(z_i, z_j)}{p(z_i)p(z_j)} \quad (2)$$

ここで、 z は画像中の word を表し、 i 番目の word、 j 番目の word の相互情報量が計算される。この相互情報量が最も値の大きいペアを Chow Liu trees にする。こうして得られる各 word の繋がりを加味して計算を行う事により、人が写ってしまう場合といった外乱などに影響されない認識を可能とする。

4. センサ統合

4-1 動作モデルによる誤認識の除去

正確な自己位置推定を可能にするために、場所認識の誤認識を取り除く事も行う。場所認識は、似た画像が写りこむと、誤認識を引き起こす可能性がある。そこで本手法では、式 (3) に示す Normalized Innovation Squared (NIS) test を導入し、誤認識を取り除く。

$$(\bar{x}_t - z_t)P_{xy}^{-1}(\bar{x}_t - z_t)^t < \delta \quad (3)$$

(a) (b)
Fig.2 Place Recognition results

ここで、 P_{xy} は残差行列、 z_t は場所認識からの観測値、 $\bar{x}_t$ は更新を行う前の状態ベクトル、 δ_t はしきい値である。このしきい値により誤認識の判断を行う。詳細は、文献 [11] を参照されたい。

4.2 Divided Difference Filter によるセンサ統合

本システムでセンサ統合に用いる Divided Difference Filter (DDF) は、非線形なモデルや観測値に、スターリン補間を用いる事で線形化の近似を行う。スターリン補間では、差分商を用いて任意の非線形関数の勾配を求める。また、採用するモデルの非線形性が強い場合であっても、線形化が安定して行える。さらに、他の手法に比べて実装が簡単だという利点もある。最近の研究において、DDF は GPS 等の観測値が不連続、不安定な状況下においても他の手法に比べて高精度な自己位置推定を行う事ができる事が知られており、今後広く利用される事が期待されている [10]。

本システムでは、二つの観測値をセンサ統合に用いる。一つは、DDF の入力値として用いるデッドレコニングである。デッドレコニングは、ホイールオドメトリのみを用いる。二つ目は、FAB-MAP の (Observation

Fig.3 Confusion matrix

(a) (b)
Fig.4 An example of the place recognition error

likelihood) を用いた場所認識である。また、場所認識は、DDF の観測値として用いる。場所認識からの観測値は、最も類似度が高いリファレンス画像の位置情報を大域的な位置として用いる。

5. 実験

本システムの有効性を示すために、つくばチャレンジ 2010 のコースで実験を行った。使用したカメラは、Logicool 社製の Qcam for Notebooks Pro (QCAM-200V) を用いる。また、Ground truth に用いる GPS は、Hemisphere 社製の A-100 を用いる。

5.1 vocabulary の算出

大量の画像から visual word を学習するために、大学校内やつくば中央公園などで取得した画像列約 1500 枚を用いて、visual words の作成を行う。集めた約 1500 枚の画像から、約 600000 もの SURF 特徴量ベクトルを取得し、k-means 法によりクラス数 10000 としてクラスタリングを行った。各クラスタの平均値を求め、代表ベクトルとして表し、代表ベクトルを 1 つの word として 10000 語の vocabulary の作成を行う。

5.2 場所認識

実験には、ロボットの進行方向左側に向けたカメラを使用し、約 2m 間隔で大学校内の画像の取得を行った。

結果を図 2 に示す。一番上に示す画像は取得した画像で、中央の図はリファレンス画像の中で、最も類似度が高い画像である。また、一番下に示すグラフは、リファレンス画像との類似度のグラフである。横軸に比較する画像の枚数を表し、縦軸には類似度を表す。左図は、ガラスの写り方が変化している例を示す。また、右図のようにわずかに人が写ってしまっている場合でも、同じ場所で取得した画像が最も類似度が高く、場所認識が行われている事がわかる。

図 3 は、走行中に 2m 間隔で取得した画像に対して、最も類似度が高いとされたリファレンス画像の対応を表したものである。横軸が問い合わせ画像、縦軸はリファレンス画像である。同じ地点から走り始め、同じ経路を走行しているため、対角線上に並ぶ点は正しく認識したとみなせる。場所認識の認識率は、73.4%で

Fig.5 Experimental results. The red line is the trajectory of our proposed method, the black line is dead reckoning, and blue marks are GPS observations.

あった。しかし似ている建物や草むらといった景色が続く実験環境のため、誤認識も生じた。誤認識した一例を図3に示す。上に示す画像は取得した画像で、下の画像は、リファレンス画像の中で最も類似度が高い画像である。また、図3に図4(a), 図4(b)の位置を示す。図4(a)や、図4(b)では、似ている草むらやフェンスといった景色が長時間続き、誤認識をしている。これらの地点では、図3で示すように場所認識からの観測値が長時間得られなかった。

5.3 自己位置推定

図5には、本手法による自己位置推定の結果を示す。この結果から、画像情報と内界センサのみで、正確な自己位置推定が行える事がわかる。また、草むらの近くで長時間の場所認識からの観測値が得られない場合でも、本システムでは正確な認識が行えている。

図6に、精度の良いGPS観測値と比較した位置誤差を示す。この図から、走行中にGPSを使わずとも、誤差が蓄積しないで、収束している事がわかる。また、今回の実験における、位置誤差の平均は24.2mであった。この結果から、本システムの有効性を示す事ができた。

6. 結言

本論文では、画像情報のみを利用した自己位置認識法について述べた。従来の自己位置推定の問題点を解決する手法について述べ、実際の画像列のデータでの実験結果を示し、提案手法の有効性を示した。

- [1] M. Cummins, P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance", *The International Journal of Robotics Research*(2008), 27 (6) 647-665
- [2] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features", *ECCV(2006)*, Graz, Austria, Vol. 13, pp. 404-417.

Fig.6 A comparison of position error

- [3] C. Chow, and C. Liu. "Approximating discrete probability distributions with dependence trees" *IEEE Transaction on Information Theory* (1968), 14(3): 462-467.
- [4] David G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", *International Journal of Computer Vision*, vol. 60, 2004, pp. 91-110
- [5] J. Sivic and A. Zisserman "Video Google: A text retrieval approach to object matching in videos" *ICCV*, Nice, France, October 2003
- [6] D. Nister and H. Stewenius "Scalable recognition with a vocabulary tree" *CVPR* volume 2, pages 2161-2168, 2006
- [7] K. Konolige, J. Bowman, JD Chen, P. Mihelich, M. Calonder, V. Lepetit, P. Fua "View-Based Maps" *In RSS'09*, 2009.
- [8] M. Tomono, "3D Localization Using Visual Odometry and Scalable Landmark-Image Retrieval", *15-th Robotics Symposia*, pp.27-32, Japan, 2010.
- [9] A. N. Ndjeng, D. Gruyer, A. Lamber, B. Mourllion, S. Glaser, "Experimental Comparison of Bayesian Outdoor Vehicle Localization Filters", *ICRA2009*, Kobe, Japan, May, 2009.
- [10] Alexandre N. Ndjeng, Dominique Gruyer, Alain Lambert, Benjamin Mourllion, Sebastien Glaser "Experimental Comparison of Bayesian Outdoor Vehicle Localization Filters", *2009 IEEE International Conference on Robotics and Automation (ICRA 2009)*, Kobe, Japan, May 12-17, 2009.
- [11] Yoichi Morales, Alexander Carballo, Eijiro Takeuchi, Atsushi Aburadani, and Takashi Tsubouchi, "Autonomous Robot Navigation in Outdoor Cluttered Pedestrian Walkways", *Journal of Field Robotics* archive Volume 26, Issue 8, 609-635, August, 2009.
- [12] <http://www.google.co.jp/help/maps/streetview/>