CH-007

2次元地図を用いた画像領域と地理情報の対応付け
Associating image regions with geographic information using 2D map

小川 将範*　　池畑 諭†　　相澤 清晴*
Masanori Ogawa　Satoshi Ikehata　Kiyoharu Aizawa

1. まえがき

市街画像において各画像上の領域毎に何の建物があるかという情報には大きな需要がある．各建物の位置が分かれば，その位置に案内のためのアノテーションを付与したり，広告を付与して対象の建物を目立たせたりといった，各建物に紐づけられた地理情報を画像中に埋め込むことができる．

この画像上の領域毎に何の建物があるかという情報を自動的に付与するシステムを構築したい．システムの入力としてはカメラ姿勢の付与された画像と地図を想定する．画像はカメラがおおよそ水平な状態かつ建物が撮影範囲に含まれる高さで撮影されていることを仮定する．画像に付与されているカメラ姿勢とは，その絶対的な位置（緯度・経度）と方向である．地図には建物の情報が含まれており，個々の建物は識別可能とする．その入手・利用が簡易であることを理由として，各建物が水平面上での頂点集合から成る多角形で表現される 2 次元地図を特に対象とする．2 次元地図には建物の高さの情報が含まれないものの，カメラの内部パラメータを既知としたときカメラ姿勢から画像の各列に対してどの建物が含まれるのかを投影モデルより計算できる．

画像に付与されているカメラ姿勢の特に位置に関しては，グローバル・ポジショニング・システム（GPS）で取得されることが多い．しかし GPS の測定結果は無視できないレベルの誤差が含まれることが知られている．そのため慣性計測装置（IMU）をはじめとしたその他センサーと組み合わせるような手法 [10, 6, 8] が存在する．それでも例えば一般的なスマートフォンでは誤差の中央値が 5m 以上であると言われる [14]．また位置情報付きの画像を提供しているサービスとして Google Street View（GSV）が有名であり，GPS と各種のセンサーを組みわせた計測を行っている [1] が，その精度は画像領域の情報を用いたアプリケーションを想定した時に十分とは言えない．

そこで本研究では，画像から推定される建物の水平方向の深度と 2 次元地図から計算される深度が近づくようなカメラ姿勢を推定する手法を提案する．

2. 関連研究

画像に対してその撮影された位置を求めるローカライゼーションの手法はコンピュータビジョンの分野において古くから研究されてきた [12]．特に，あらかじめ位置情報が付与された画像データベースを参照するものと，建物などの形状情報のみが含まれる地図を参照する手法が存在し，本手法と関係の深い後者について詳しく述べる．

地図を参照する手法では，入力画像から注目するオブジェクトの形状情報を抽出し，地図上でそれに対応するような地点を探索する．地図上で対象のオブジェクトが 3 次元的に表現されている場合の手法には [2, 11] がある．今回の提案手法と同様に地図上で建物が 2 次元の頂点集合で表されていることを想定する手法にもいくつかの研究が存在する．[3] は，画像中から検出された直線の中から消失点との関係を元に鉛直な角であるものを抽出し，その直線に接している面の法線から建物を上から見た時の輪郭片の集合を計算し，地図上でそれらに対応するようなカメラ位置と方向を推定する．[5] はおおよそ中央に建物が 1 つだけ存在するような画像を前提としているものの，[3] の手法をカメラの傾きを消失点に基づいて計算しその値に応じた正規化を行うことで，精度の向上を実現している．[13] は，カメラの方向を既知としたとき初期値周辺のカメラ位置について地図上の建物の頂点を画像中に投影した部分が，画像上でどの程度建物のかどの鉛直な直線らしいかを集計し，全ての地図上の頂点を投影した時の合計がもっとも高いカメラ位置を推定値としている．[9] は画像中から建物の角である鉛直な直線を推定する際に，セマンティックセグメンテーションを組み合わせることでその信頼度を向上させている．

2 次元地図を参照して画像の撮影されたカメラ位置をする手法 [3, 5, 13, 9] では，建物の角である鉛直な直線を画像から抽出し，それらと地図上の建物の頂点を対応づけようとする点で共通しているため，それらが建物以外のオブジェクトによって隠されてしまうと必要な情報の多くを失ってしまう．それに対し，本論文で提案する手法では，建物以外のオブジェクトによって建物の一部分が隠されてしまうような状況においても，残りの見えている部分を利用できるようなアプローチを取っている．

3. 提案手法

本手法では 2 次元地図とあるカメラ姿勢が与えられたときに，建物までの水平方向の深度が計算できることを利用する．画像から推定する建物領域の水平方向の深度と地図の投影から得られる深度が近くなるように，画像に付与されているカメラ姿勢の周辺において探索することで，カメラ姿勢を推定する．その概要を図 1 に示す．その後，得られたカメラ姿勢を元に地図中の建物を画像の建物領域に投影することで，各建物を画像中のそれぞれ領域に対応づける．以下にその詳細を記述する．また本手法では距離計算を単純にするため，水平面上の位置である緯度経度をユニバーサル横メルカトル座標に変換して用いる．

*東京大学大学院 情報理工学系研究科 電子情報学専攻
†国立情報学研究所

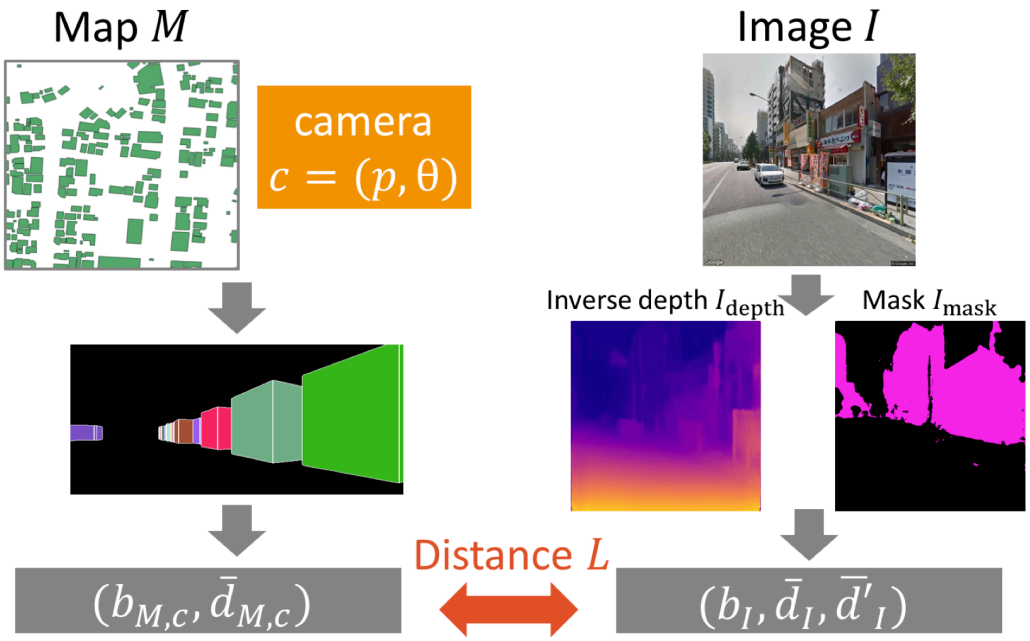

図 1: 建物の水平深度を用いたカメラ姿勢推定の概要

3.1. 手法の入出力

本手法は入力として 2 次元地図と位置情報のついた市街環境で撮影された画像を用いる．2 次元地図 M は建物の集合で構成されており，それぞれの建物は水平面上の位置を表す頂点から成る多角形として表現される．地図中の建物の数を N_{building} として，どの建物かを表す識別子の集合 $\{1, 2, ..., N_{\mathrm{building}}\}$ と建物でないことを表す識別子 0 を合わせた集合を S_{building} とする．また画像 I は，カメラがおおよそ地面に対して水平な状態かつ建築物が画角の範囲に入る程度の高さで撮影されており，カメラの内部パラメータは既知であると仮定する．画像に付与されているカメラ姿勢（位置・方向）を c_{init} とする．

また手法の出力は，画像 I と同じ大きさで建物の識別子を $R(i, j) \in S_{\mathrm{building}}$ で与える R である（ここでは i が列方向，j が行方向のインデックスであり，$i = 1, 2, ..., N_{\mathrm{col}}$ とする）．

3.2.2 次元地図とカメラ姿勢からの水平深度の計算

仮定よりカメラの内部パラメータが既知でありカメラの回転が鉛直方向を軸とした 1 自由度に限定される時，2 次元地図 M に対してあるカメラ姿勢 c から透視投影モデルに基づいて水平方向の 1 次元の深度系列が決定される．この深度系列を N_{col} 次元ベクトルで表現したものを $D_{M,c}$ とする．同時に各位置で対応する建物がどれであるかを表す $B_{M,c}$ も決定される．建物が存在しない，すなわち $B_{M,c}(n) = 0$ であるような位置 i では深度は無限となる（$D_{M,c}(i) = \infty$）．

深度が大きい時はその小さな変化によって観測される画像はあまり変化しないが，深度が小さいときはその変化が小さくても観測される画像に大きな変化を及ぼす．そのため幅 N_{col} の深度の逆数の系列を $d_{M,c}$ とし，それの各要素が $[0, 1]$ の範囲に含まれるように正規化したものを $\bar{d}$ とすると，

$$d_{M,c}(i) = \frac{\mathrm{const}}{D_{M,c}(i)},$$
$$\bar{d}_{M,c}(i) = \frac{d_{M,c}(i)}{\max_{i'}(d_{M,c}(i'))}$$

となる．また各位置において建物の存在の有無を表す系列 $b_{M,c}$ は

$$b_{M,c}(i) = \begin{cases} 1 & (B_{M,c}(i) \neq 0) \\ 0 & (B_{M,c}(i) = 0) \end{cases}$$

となる．この $\bar{d}_{M,c}$ と $b_{M,c}$ を地図 M 上でカメラ c が観測できる周辺環境を表現する特徴量として利用する．

3.3. 画像からの建物領域の水平深度の推定

画像 I が与えられた時に，その水平方向の様子を正規化深度の逆数の系列 $\bar{d}_I$ と建物の有無 b_I で表現したい．そこで画像 I に対し，各画素が建物か否かを表した I_{mask} と，各画素での深度の逆数 I_{depth} を推定する．

I_{mask} はセマンティックセグメンテーションの手法を用いて，その結果から建物領域と推定された部分のみを 1，それ以外を 0 とする．I_{depth} は画像からの深度推定手法を用いる．セマンティックセグメンテーションや深度推定手法には数多くの研究がなどが挙げられるが，今回はそれぞれ [4]，[7] を用いた．

次に推定された I_{mask} と I_{depth} から水平方向の深度の逆数の系列 d_I と b_I を計算する．地図 M 上には建物しか存在しないためその位置から建物までの深度を計算できたが，実際の環境には建物以外のオブジェクトが存在するため I_{depth} は必ずしも建物までの深度ではない．そこで b_i は列中に建物領域が存在するか否かで判別し，列中の建物領域の中で最大の深度の逆数を d_I

とし，建物かどうかにかかわらず列中の最大の深度の逆数を d'_I とする．これは地図から計算された深度系列との距離を定義する際に，建物の前にオブジェクトが存在することによってその後ろ側の建物の深度の情報が得られない列に関して無視をするために用いる．画像中の位置を列と行の組み (i,j) とした時，その計算は以下のようになる．

$$b_I(i) = \max_j I_{\mathrm{mask}}(i,j),$$
$$d_I(i) = \max_j (I_{\mathrm{mask}} \odot I_{\mathrm{depth}})(i,j),$$
$$d'_I(i) = \max_j I_{\mathrm{depth}}(i,j).$$

また d_I, d'_I の各要素を d_I の最大値で割ることによって，$\bar{d}_I, \bar{d'}_I$ が計算される．

3.4. 水平深度系列間の距離定義

上記の手順によって周辺環境を表現する特徴量について地図とカメラパラメータから $(\bar{d}_{M,c}, b_{M,c})$ が，画像から $(\bar{d}_I, \bar{d'}_I, b_I)$ がそれぞれ得られた．これらの間の距離尺度 L を定義し，それを最小化するようなカメラ姿勢を推定したい．

理想的に建物までの深度が推定されている場合，各位置での深度の逆数の差の平均をとることで特徴量の距離としたいが，画像中には建物以外のオブジェクトが存在し，そのオブジェクトによってカメラから建物が隠されてしまう場合が存在する．あるカメラパラメータが画像での観測状態を再現できているときに，ある位置において地図を投影した結果では建物が存在していながら画像では観測されておらず，その位置での地図からの建物の深度の逆数が画像からのものよりも小さい時，すなわち

$$S_{\mathrm{hide}} = \{i | b_{M,c}(i) = 1, b_I(i) = 0, \bar{d}_{M,c}(i) < \bar{d'}_I(i)\}$$

に含まれる位置では，何らかのオブジェクトによって建物までの深度が観測できない可能性がある．このような位置以外において各位置での建物までの深度の逆数の差の 2 乗の平均を，地図からの特徴量と画像からの特徴量の距離尺度 L とすると，

$$L = \frac{1}{N_{\mathrm{col}} - |S_{\mathrm{hide}}|} \sum_{i \in \{1,2,\ldots,N_{\mathrm{cols}}\} \setminus S_{\mathrm{hide}}} (\bar{d}_{M,c}(i) < \bar{d}_I(i))^2$$

となる．

入力として地図 M と画像 I が固定されたとき，L はカメラ姿勢 c の関数と見ることができる．画像に付与されているカメラ姿勢 c_{init} の周辺において

$$\hat{c} = \arg\min_c L(c)$$

となるような $\hat{c}$ をカメラ姿勢の推定値とする．

3.5. 推定されたカメラ姿勢に基づく建物の投影

推定されたカメラ姿勢 $\hat{c}$ を用いて，画像中の建物領域範囲内に建物の識別子を投影することで，各建物が画像中のどの領域に対応するのかを求める．位置 (i,j) に対して対応する建物の識別子は

$$R(i,j) = B_{M,\hat{c}}(i) \cdot I_{\mathrm{mask}}(i,j)$$

で与えられる．

カメラ姿勢	大島町	文京区	新宿区	全体
初期値	0.779	0.769	0.799	0.785
推定地	0.808	0.812	0.845	0.827

表 1: 建物の識別子の正解率

4. 実験

提案する地図と画像から得られる水平深度系列間距離の最小化によるカメラ姿勢推定が，各建物の領域推定に有効であるかを検証した．

4.1. 実験条件

実験を行う対象として，2 次元地図とカメラ姿勢の付与された画像群を用意した．2 次元地図としては 3 地域（東京都の大島町，文京区，新宿区）の NTT 空間情報株式会社の GEOSPACE 電子地図データから家屋レイヤを用いた．カメラ姿勢の付与された画像群は Google の Street View API を用いて収集した．東京都の大島町，文京区，新宿区の範囲内からそれぞれ 20, 30, 30 の建物が画像内に含まれるような異なるカメラ姿勢をクエリとし，画像の大きさは縦横 640 ピクセル，水平画角は 90 度である．

今回の実験では推定対象は誤差が生じやすいカメラ位置とし，カメラ方向は付与されているものをそのまま用いた．付与されているカメラ位置を (x_0, y_0) としたとき，カメラ位置を $(x_0 + \Delta x, y_0 + \Delta y)$ として $\Delta x, \Delta y$ をそれぞれ $[-3\mathrm{m}, 3\mathrm{m}]$ から 0.1m 毎にサンプルした点を探索範囲とした．

また，評価のために各画像において地図上の建物の投影結果が画像中の建物と一致するようなカメラ位置を人手で探索し，このときのカメラ姿勢を c_{GT} とした．

4.2. 評価

提案手法によって推定されたカメラ姿勢 $\hat{c}$ に基づいた地図の投影結果が，あらかじめ画像に付与されていた Google Street View によるカメラ姿勢 c_{init} に基づいたものと比較し，画像中の建物とどの程度一致しているかを評価した．各ピクセルにおいて，その建物の識別子が c_{GT} に基づく投影結果と一致しているときに正解とし，対象となる画像群の推定された建物領域の範囲内での正解率を表 1 に示す．これより全体においても，各地域においても提案手法によって正解率が数パーセント向上していることが確認できる．また，図 2，図 3 にそれぞれ提案手法によって正解率が向上した画像と低下した画像の例を示す．各地点において左から順にそれぞれ $c_{\mathrm{GT}}, c_{\mathrm{init}}, \hat{c}$ に基づいて投影した結果であり，マスクの色が建物の識別子を表している．傾向として図 2 のように画像の水平方向に深度が大きく変化している場合，手法が上手くいきやすい．反対に図 3 の (a) のように画像面と平行な壁が全体に写っていたり，(b) のように建物があったとしてもカメラから遠いなどの，深度が全体として一様でカメラ位置の移動によってあまり変化しないと考えられる場合に手法が失敗しやすい．

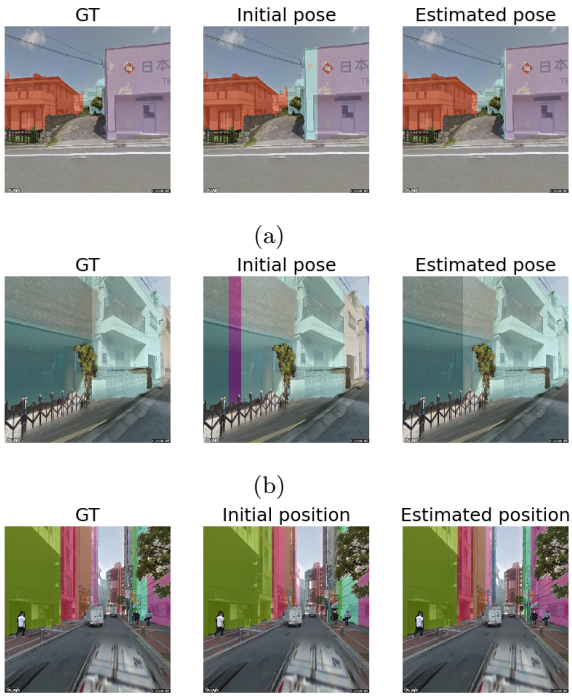

図 2: 提案手法の効果が確認されたケース

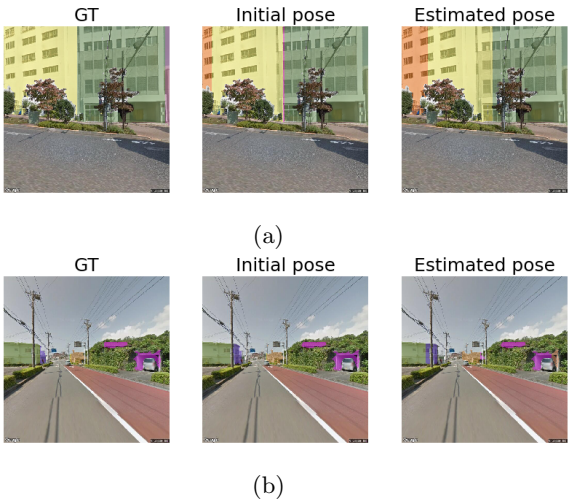

図 3: 提案手法が失敗したケース

5. まとめ

地図中の個々の建物を画像中のそれぞれの領域に対応づけることを目的として，2 次元地図とあるカメラ姿勢が与えられたときに，建物の水平方向の深度が計算できることを利用し，画像から推定された建物領域の深度に地図の投影から得られる深度が近くなるようなカメラ姿勢を推定する手法を提案した．Google Street View に付与されているカメラ姿勢の精度ですら投影によって建物を識別するには不十分であり，提案手法によって個々の建物の領域推定の精度が向上することが確認できた．

今後の展望として，画像のピクセル毎に建物のラベルを付けたデータセットを用意することでより正確な評価をできるようにすると共に，既存手法との比較を行なっていきたい．

謝辞

本研究の一部は，VTEC 研究所からの支援を受けた．

参考文献

[1] Dragomir Anguelov, Carole Dulong, Daniel Filip, Christian Frueh, Stéphane Lafon, Richard Lyon, Abhijit Ogale, Luc Vincent, and Josh Weaver. Google street view: Capturing the world at street level. *Computer*, Vol. 43, No. 6, pp. 32–38, 2010.

[2] Mayank Bansal and Kostas Daniilidis. Geometric urban geo-localization. In *CVPR*, pp. 3978–3985. IEEE, 2014.

[3] Tat-Jen Cham, Arridhana Ciptadi, Wei-Chian Tan, Minh-Tri Pham, and Liang-Tien Chia. Estimating camera pose from a single urban ground-view omnidirectional image and a 2d building outline map. In *CVPR*, pp. 366–373. IEEE, 2010.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018.

[5] Hang Chu, Andrew Gallagher, and Tsuhan Chen. Gps refinement and camera orientation estimation from a single image and a 2d map. In *CVPR Workshops*, pp. 171–178. IEEE, 2014.

[6] Nabil Mohamed Drawil and Otman Basir. Intervehicle-communication-assisted localization. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 11, No. 3, pp. 678–691, 2010.

[7] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, Vol. 2, p. 7, 2017.

[8] Kichun Jo, Keounyup Chu, and Myoungho Sunwoo. Interacting multiple model filter-based sensor fusion of gps with in-vehicle sensors for real-time vehicle positioning. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 13, No. 1, pp. 329–343, 2012.

[9] Arsalan Mousavian and Jana Kosecka. Semantic image based geolocation given a map. *arXiv preprint arXiv:1609.00278*, 2016.

[10] Honghui Qi and John B Moore. Direct kalman filtering approach for gps/ins integration. *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 38, No. 2, pp. 687–693, 2002.

[11] Olivier Saurer, Georges Baatz, Kevin Köser, Marc Pollefeys, et al. Image based geo-localization in the alps. *International Journal of Computer Vision*, Vol. 116, No. 3, pp. 213–225, 2016.

[12] Richard Szeliski. Where am i?" : Iccv 2005 computer vision contest, 2005.

[13] Jiangye Yuan and Anil M Cheriyadat. Combining maps and street level images for building height and facade estimation. In *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics*, p. 8. ACM, 2016.

[14] Paul A Zandbergen and Sean J Barbeau. Positional accuracy of assisted gps data from high-sensitivity gps-enabled mobile phones. *The Journal of Navigation*, Vol. 64, No. 3, pp. 381–399, 2011.