Vol.2015-CVIM-195 No.60
2015/1/23

特徴点の明示的な対応付けを伴わないカメラ位置姿勢推定

黒川 陽平[1,a)]　中島 悠太[1]　佐藤 智和[1]　横矢 直和[1]

概要：モバイル端末での利用を想定した軽量なカメラ位置姿勢推定の実現に向けて，本稿では類似度による特徴点の明示的な対応付けを伴わない手法について検討する．提案手法では，事前に生成した 3 次元点群データベースに含まれる特徴点を対象画像上に投影して得られる投影点と対象画像から得られる特徴点の画像上での位置の一致度合いを評価尺度とし，これに基づいてカメラの位置姿勢を推定する．

キーワード：カメラ位置姿勢推定，Structure from Motion，3 次元点群，自然特徴点，拡張現実感

1. はじめに

モバイル端末で動作する拡張現実感技術を用いたアプリケーションでは，端末のカメラで撮影した実世界の動画像に対し視覚情報を重畳して提示するために，カメラの位置姿勢を高速・高精度に推定する必要がある．拡張現実感のためのカメラ位置姿勢推定手法として，マーカを用いる手法 [1] が広く利用されているが，環境中にマーカを配置する必要があるため，利用可能な環境が限定される．

マーカを用いずに高精度なカメラ位置姿勢推定を実現する手法として，事前に構築した 3 次元点群データベースを用いる手法が提案されている [2], [3], [4]．これらの手法は，カメラから得られる画像（カメラ画像）から検出した特徴点と，点群データベースに含まれる 3 次元点を対応付け，Perspective-n-Point 問題を解くことによってカメラ位置姿勢を推定する．特徴点と 3 次元点の対応付けには点の周辺の見えに関する記述子が用いられる．

この手法は，高精度な推定が可能である一方で，カメラ画像からの記述子抽出や記述子による点の対応付けの計算コストが高く，実時間性が要求される拡張現実感への応用は難しい．さらに，点群データベースのすべての 3 次元点について対応付けのための記述子を保持する必要があるため，データベースのサイズが肥大化しやすく，モバイル端末での取り扱いが困難である．

そこで著者らは，記述子による特徴点の明示的な対応付けを伴わず，3 次元点の座標のみを持つ 3 次元点群データベースを利用したカメラ位置姿勢推定手法提案した [5]．この手法では，カメラ画像中の特徴点の空間的な分布，および 3 次元点群データベースに含まれる点群の分布の偏りに着目してカメラ位置姿勢を推定する．すなわち，任意のカメラ位置姿勢のもとでデータベース中の点群を画像平面に投影した際の点の分布と，クエリ画像中の特徴点分布の一致度合いを評価する尺度を定義し，カメラ位置姿勢に関してこの評価尺度を最適化することにより位置姿勢を求める．この手法は，明示的な点の対応付けが不要となる一方で，たとえばカメラの位置を無限遠に設定してデータベース中のすべての点を単一の特徴点に投影させると評価尺度が大きくなるため，位置姿勢推定の誤差が増大するという問題があった．

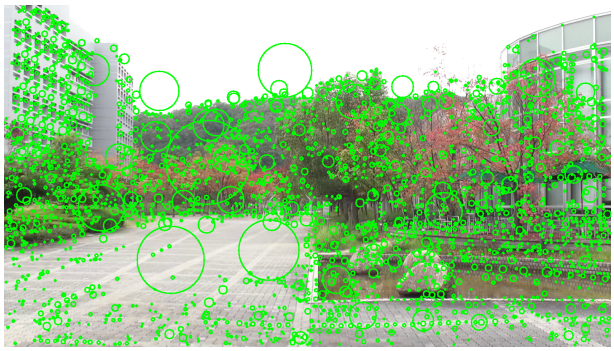


図 1: カメラ画像と SIFT 特徴点（緑の円）の例．円の中心が特徴点の位置を，円の大きさが特徴点のスケールを示す．

そこで本報告では，カメラ画像上の特徴点検出に用いられる SIFT 特徴点 [6] のスケールを考慮した手法を提案する（図 1）．SIFT 特徴点の検出はブロブ検出とみなすことができ，このときスケールは検出されるブロブのカメラ画像上での大きさに対応する．したがって，カメラの位置姿勢，および内部パラメータを利用することにより，スケールから 3 次元空間中でのブロブに対応する構造の大きさを

[1] 奈良先端科学技術大学院大学
[a)] kurokawa.yohei.kn3@is.naist.jp

© 2015 Information Processing Society of Japan

推定可能である．本報告ではこの大きさを 3 次元点スケールと呼ぶ．提案手法では，3 次元点群データベースの構築において，各点に対して画像上のスケールに基づく 3 次元点スケールを付与する．位置姿勢推定では，3 次元点群を画像に投影する際に，同様にして図 2 に示すように 3 次元点スケールを投影することにより，その 3 次元点が観測されるスケールが求められることを利用して，SIFT 特徴点検出のスケールごとにカメラ位置姿勢の評価尺度を定義し，これらを統合して最終的な評価尺度を求める．これにより，それぞれのスケールにおける評価尺度の自明な極大値の影響を抑えた位置姿勢推定を可能とする．

2. 特徴点の明示的な対応付けを伴わないカメラ位置姿勢推定

提案手法は，3 次元点スケールを含む 3 次元点群データベースの構築，およびリアルタイムで取得されるカメラ画像に対するカメラの位置 (t_x, t_y, t_z) と姿勢 (r_x, r_y, r_z) の計 6 パラメータ $\boldsymbol{\theta} = (r_x, r_y, r_z, t_x, t_y, t_z)$ の推定，から構成される．ここではこれらについて詳述する．

2.1 3 次元点群データベース構築

3 次元点群データベースは，点群 $P = \{\mathbf{p}_j | j = 1, \ldots, J\}$ と，それぞれの点 $\mathbf{p}_j = (x_j, y_j, z_j)$ に対応する 3 次元点スケール S_j から構成される．点群 P は，事前に撮影された対象環境の画像群を入力として，Structure from Motion (SFM) 法 [7] により与えられる．SFM 法は，多くの場合，入力画像群中の画像 $I_n (n = 1, \ldots, N)$ から SIFT 特徴点とそれぞれの特徴点に対する記述子を算出し，これらに基づいて 3 次元点群 P と入力画像群のカメラ位置姿勢を求める．このとき，点 $\mathbf{p}_j$ が検出された入力画像 I_n 中の SIFT 特徴点のスケール $s_{n,j}$ は，対応する 3 次元形状やテクスチャを点 $\mathbf{p}_j$ を中心とする球によって表したときの半径に比例すると考えられる．一般に，焦点距離 f のカメラから奥行き z の位置にある半径 S の球が，画像上で半径 s の円として観測された場合，その関係は $S = zs/f$ で与えられる．点 $\mathbf{p}_j$ は複数の入力画像で観測されることから，提案手法では点 $\mathbf{p}_j$ に対する 3 次元点スケール S_j を，複数画像のスケールに基づく値の平均として，次式により定義する．

$$S_j = \frac{1}{N_j} \sum_n \frac{z_{n,j} s_{n,j}}{f_n} \tag{1}$$

ここで，$z_{n,j}$ は画像 I_n を撮影したカメラ中心から $\mathbf{p}_j$ までの奥行き，f_n は画像 I_n を撮影したカメラの焦点距離，N_j は $\mathbf{p}_j$ を観測した入力画像数であり，和はこれらの画像に対して計算される．得られた 3 次元点群 P と対応する 3 次元点スケールをデータベースに保存する．

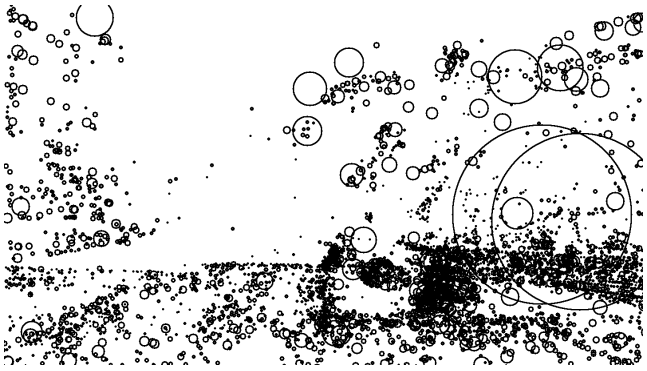


図 2: 3 次元点スケールの投影の例．

2.2 点の分布に関する評価尺度

図 1 に示すように，カメラ画像から得られる特徴点の分布には，撮影した環境の 3 次元形状やテクスチャによる偏りが生じる．また，同様に特徴点を利用して得られるデータベース中の 3 次元点群にも偏りが生じる．提案手法ではこの偏りに着目し，特徴点の分布，および 3 次元点群を画像に投影して得られる投影点の分布の一致度合いに関する評価尺度を定義して，その尺度をカメラ位置姿勢に関して最大化することにより，位置姿勢を推定する．

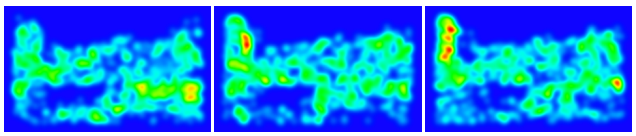
提案手法はまず，リアルタイムに得られるカメラ画像から SIFT 特徴点を検出する．得られた特徴点集合を $Q = \{\mathbf{q}_i | i = 1, \ldots, M\}$ とし，それぞれの特徴点 $\mathbf{q}_i = (u_i, v_i)$ にスケール s_i が与えられる．SIFT 特徴点検出におけるスケールは離散的な値をとることから，$s_i \in \{\sigma_k | k = 1, \ldots, K\}$ とし，まずスケール毎の点の分布を考える．ここでは，スケール σ_k における点の分布 $d_k(\mathbf{q})$ を，ガウシアンカーネル $g(\mathbf{q}) = \exp(-\beta \mathbf{q}^\top \mathbf{q})$ により次式で表す．

$$d_k(\mathbf{q}) = \frac{1}{|\Phi_k|} \sum_{i \in \Phi_k} g(\mathbf{q} - \mathbf{q}_i) \tag{2}$$

ただし，Φ_k はスケール σ_k で検出された特徴点の添え字集合を表す．図 3 に示すように，点の分布 $d_k(\mathbf{q})$ は対象環境の 3 次元形状，およびテクスチャに依存するものであり，点の密度が高いほど大きな値となる．

任意のカメラ位置姿勢 $\boldsymbol{\theta}$ が与えられたときのデータベース中の 3 次元点群の投影点と，カメラ画像から検出される特徴点の分布の一致度を評価する尺度 ρ は，$d_k(\mathbf{q})$ を利用して次式により定義する．

$$\rho(\boldsymbol{\theta}) = \sum_{j=1}^{J} d_{k_j}(\mathrm{proj}_{\boldsymbol{\theta}}(\mathbf{p}_j)) \tag{3}$$



(a) σ_1 (b) σ_2 (c) σ_3

図 3: 図 1 に示す特徴点の分布 $d_k(\mathbf{q})$ の例．

© 2015 Information Processing Society of Japan

ただし，$\mathrm{proj}_{\boldsymbol{\theta}}(\mathbf{p}_j)$ は，既知のカメラ内部パラメータとカメラ位置姿勢 $\boldsymbol{\theta}$ による 3 次元点 $\mathbf{p}_j$ の画像平面への投影を表す．また，k_j は $\mathbf{p}_j$ に対応する 3 次元点スケール S_j をカメラ内部パラメータと位置姿勢 $\boldsymbol{\theta}$ により画像平面に投影したときのスケールの添え字であり，S_j を画像平面に投影したときの大きさ $S_j f/z_j$ が σ_{k_j} にもっとも近くなるように選ばれる．式 (3) により定義される評価尺度 ρ は，特徴点と $\boldsymbol{\theta}$ による投影点の分布の一致度合いが高いとき大きな値となる．図 3 に示すように，各スケールにおいて $d_k(\mathbf{q})$ の極大値の位置が異なることから，スケールを利用することで，スケールを利用しない手法 [5] において無限遠のカメラ位置で評価尺度が大きな値をとる問題が緩和されるものと考える．

2.3 ルックアップテーブルによる評価尺度算出の高速化

提案手法では，式 (3) を最大とする $\boldsymbol{\theta}$ を求めることで，カメラの位置姿勢を推定する．しかし，式 (3) は異なる $\boldsymbol{\theta}$ での評価尺度の算出に MJ 回のガウシアンカーネルの計算が必要となり，リアルタイム性が要求されるモバイル端末上での利用には計算コストが高い．そこで，$d_k(\mathbf{q})$ をルックアップテーブルとして保持する．$d_k(\mathbf{q})$ は，特徴点を 1 画素の点としてプロットした画像にガウシアンカーネルを畳み込んだものとして近似できる．ルックアップテーブルの生成はカメラ画像を取得するごとに 1 回で，これにより J 回のテーブル参照によって評価尺度の近似値が得られる．

2.4 評価尺度の最大化

前述のように，提案手法では式 (3) の最大化によりカメラ位置姿勢 $\boldsymbol{\theta}$ を推定する．しかし，式 (3) はガウシアンカーネルを利用することから，図 3 に示すように多数の極値を持つため，たとえば勾配方向を利用した最適化手法では真のカメラ位置姿勢を求めることは困難であると考える．一方，提案手法はリアルタイムでの利用を想定しており，位置姿勢の連続性から連続するカメラ画像の位置姿勢は高い相関を持つ．そこで，パーティクルフィルタ [8] を利用して位置姿勢を推定する．ここでは，パーティクルフィルタにおける尤度 $L(\boldsymbol{\theta})$ として，

$$L(\boldsymbol{\theta}) = \exp(\rho(\boldsymbol{\theta})) \tag{4}$$

を用い，運動モデルとしてある時刻 τ における位置姿勢 $\boldsymbol{\theta}_\tau$ は，前時刻の位置姿勢 $\boldsymbol{\theta}_{\tau-1}$ を利用して，

$$\boldsymbol{\theta}_\tau = \boldsymbol{\theta}_{\tau-1} + \boldsymbol{\nu} \tag{5}$$

で与えられるものとする．ただし $\boldsymbol{\nu}$ は平均 $\mathbf{0}$，分散 $\eta^2\mathbf{I}$ のガウス分布に従うものとする．

3. 実験と評価

ここでは，評価尺度の有用性を示すために評価尺度の振る舞いを検証する実験と，提案手法によるカメラ位置姿勢推定の精度に関する実験を行った．実験にあたって，3 次元点群データベースの構築，およびカメラ内部パラメータ推定に VisualSFM [7] を用いた．得られた 3 次元点群（図 4）は 22,147 点で，実世界の 1 m が並進成分のおよそ 0.4 に相当する．カメラ画像からの特徴点検出には SiftGPU [9] を利用した．$d_k(\mathbf{q})$ のルックアップテーブルを生成する際の離散ガウシアンカーネルは経験的に 5×5 画素とした（$\beta = 1/1.21$ に相当）．なお，本実験では，カメラ位置姿勢 $\boldsymbol{\theta}$ の真値を点群データベース構築時に用いた特徴点との対応付けに基づいて推定した．

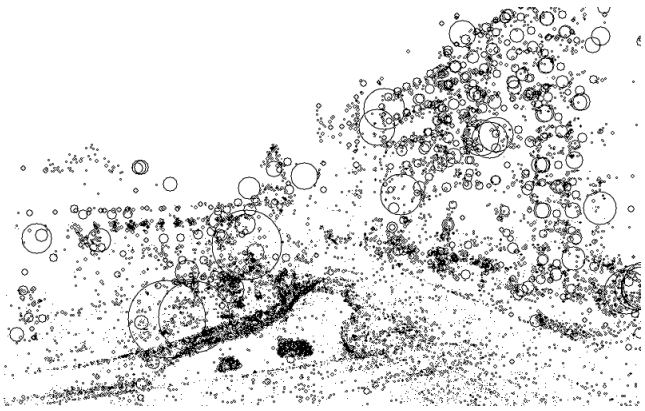
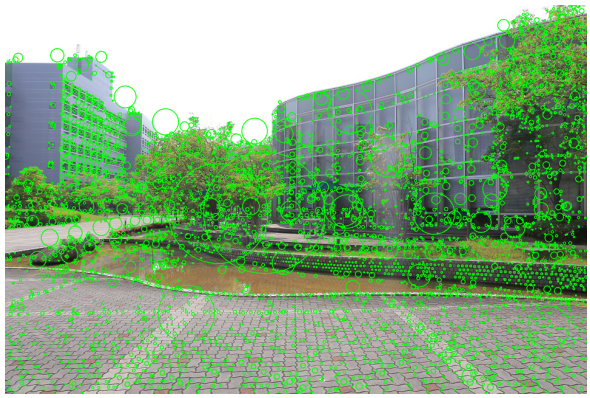


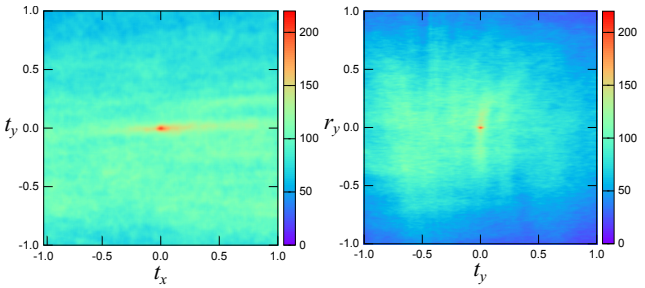
図 4: 実験に用いた 3 次元点群

3.1 評価尺度の振る舞いの検証

式 (3) として提案した評価尺度が真値において最大値と



(a) 入力画像と検出した SIFT 特徴点



(b) **(c)**

図 5: 評価尺度のカラーマップ

© 2015 Information Processing Society of Japan

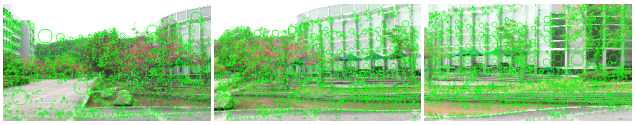


図 6: 入力フレームの例

なることを確認するために，パラメータ $\boldsymbol{\theta}$ の変化に応じて $\rho(\boldsymbol{\theta})$ がどのように振る舞うかを検証する実験を行った．実験では，6 成分からなる $\boldsymbol{\theta}$ の 4 成分を真値に固定し，残りの 2 成分を真値付近で連続的に変化させながら $\rho(\boldsymbol{\theta})$ を算出し，カラーマップとしてプロットした．

図 5(a) に入力画像と検出した SIFT 特徴点を，図 5(b)，5(c) に入力画像に対する評価尺度の振る舞いをそれぞれ示す．グラフの縦軸・横軸は，変化させた 2 成分の真値からのオフセット量であり，回転成分の単位はラジアン，並進成分は先述の通り 0.4 が実世界の 1 m に相当する．すなわち，カラーマップの中央 $(0,0)$ が真値に対応し，そこから離れるにつれて真値との誤差が増大する．これより，カラーマップの中央に高いピークがあることが見て取れ，提案した評価尺度がカメラ位置姿勢の評価に有用であることがわかる．

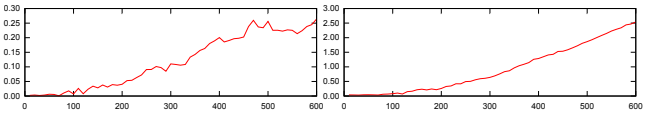
3.2 カメラ位置姿勢推定の実験・評価

動画像を入力とした提案手法によるカメラ位置姿勢推定の評価実験を行った．実験で利用した動画像は 1,440 × 810 画素で，全 600 フレームである（図 6）．フレームごとに SIFT 特徴点を検出した．ここでは，初期フレームのカメラ位置姿勢を真値とし，以降パーティクルフィルタにより推定した．式 (5) で示したカメラの運動モデルのパラメータは $\eta = 0.005$ とした．パーティクル数は 300 個で，このうち尤度最高となったパーティクルを推定値として採用した．推定した位置姿勢は 10 フレームごとに真値と比較し，真値に対する回転誤差・並進誤差を求めることで評価した．回転誤差・並進誤差はそれぞれ次式で求められる．

$$E_{\mathrm{rot}}\left(\mathbf{R}, \mathbf{R}_{\mathrm{GT}}\right) = \cos^{-1}\left(\frac{tr\left(\mathbf{R}\mathbf{R}_{\mathrm{GT}}^{-1}\right) - 1}{2}\right) \tag{6}$$

$$E_{\mathrm{tra}}\left(\mathbf{t}, \mathbf{t}_{\mathrm{GT}}\right) = \left\|\mathbf{t} - \mathbf{t}_{\mathrm{GT}}\right\|_2 \tag{7}$$

ただし，$\mathbf{R}$ および $\mathbf{t}$ はそれぞれカメラ位置姿勢パラメータ $\boldsymbol{\theta}$ から求められる回転行列と並進ベクトルであり，$\mathbf{R}_{\mathrm{GT}}$ および $\mathbf{t}_{\mathrm{GT}}$ は真値である．図 7 に実験結果を示す．グラフは 10 フレームごとの推定誤差を示しており，横軸がフレーム番号を，縦軸が誤差をそれぞれ表している．推定開始後は比較的小さな誤差で推定が可能であったものの，時間経過に伴って誤差が増大していることがわかる．誤差が増大する理由として，パーティクル数の不足，最尤パーティクルを推定値として採用することが不適などが考えられる．



(a) 回転誤差　(b) 並進誤差

図 7: カメラ位置姿勢推定の推定誤差

4. まとめ

本報告では，記述子による特徴点の明示的な対応付けを伴わないカメラ位置姿勢推定手法を提案した．また実験により，提案した評価尺度の振る舞いが位置姿勢の推定に適していることを確認した．さらに，動画像を入力とした提案手法によるカメラ位置姿勢推定の精度について確認した．結果として，パーティクルフィルタを用いた評価尺度の最大化による位置姿勢推定は，パラメータ設定などに関するさらなる検討を要するものの，推定開始後は推定誤差が小さいことから，提案手法の有用性を示すことができたと考えられる．今後の課題として，評価尺度の最適化の際に，ガウシアンカーネルの大きさを変更する粗密法的なアプローチの採用などが考えられる．

謝辞　本研究の一部は科学研究費補助金 (No. 23240024, No. 26330193) による．

参考文献

[1] Kato, H. and Billinghurst, M.: Marker Tracking and HMD Calibration for a Video-based Augmented Reality Conferencing System, *Proc. IEEE and ACM Int. Workshop on Augmented Reality*, pp. 85–94 (1999).
[2] Irschara, A., Zach, C., Frahm, J.-M. and Bischof, H.: From Structure-from-motion Point Clouds to Fast Location Recognition, *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, pp. 2599–2606 (2009).
[3] Li, Y., Snavely, N., Huttenlocher, D. and Fua, P.: Worldwide Pose Estimation Using 3D Point Clouds, *Proc. European Conference on Computer Vision*, pp. 15–29 (2012).
[4] Sattler, T., Leibe, B. and Kobbelt, L.: Fast Image-based Localization using Direct 2D-to-3D Matching, *Proc. IEEE Int. Conf. Computer Vision*, pp. 667–674 (2011).
[5] 黒川陽平，中島悠太，佐藤智和，横矢直和：特徴点の類似度尺度による対応付けを伴わないカメラ位置姿勢推定手法の検討，2014 年映像情報メディア学会年次大会講演予稿集，2 pages (2014).
[6] Lowe, D. G.: Distinctive Image Features from Scale-Invariant Keypoints, *Int. J. Computer Vision*, Vol. 60, No. 2, pp. 91–110 (2004).
[7] Wu, C.: VisualSFM : A Visual Structure from Motion System, (online), available from ⟨http://ccwu.me/vsfm/⟩ (accessed 2014-12-15).
[8] Isard, M. and Blake, A.: CONDENSATION - Conditional Density Propagation for Visual Tracking, *Int. J. Computer Vision*, Vol. 29, pp. 5–28 (1998).
[9] Wu, C.: SiftGPU: A GPU Implementation of Scale Invariant Feature Transform (SIFT), (online), available from ⟨http://cs.unc.edu/ ccwu/siftgpu/⟩ (accessed 2014-12-15).

© 2015 Information Processing Society of Japan