Vol.2017-CVIM-205 No.44
2017/1/20

固定多視点カメラ画像を利用した移動カメラの位置姿勢推定

小畑 圭 *1　　斎藤 英雄 *1

Camera Pose Estimation Using Real Image Sequence Captured by Fixed Cameras

Kei Obata*1 and Hideo Saito*1

Abstract – 移動カメラの位置姿勢推定は，それで撮影された画像中の **2** 次元位置と，空間中の **3** 次元位置の正しい対応付けによって，そのパラメータが求められる．従来手法として，**3** 次元位置の取得のために予めシーンのテクスチャ付き **3** 次元形状モデルを作製し，それと撮影された画像の間での **2D-3D** 対応に基づいてカメラの位置姿勢を推定するものがある．この手法は，モデルの正確な作製が難しい複雑な形状や表面を持つシーンである場合，適用が困難であった．そこで我々は，あらかじめシーンを撮影した実画像列と，移動カメラで撮影された画像の間での **2D-2D** 対応に基づき，そのカメラの位置姿勢を推定する手法を提案する．提案手法では，複数の **2D-2D** 対応からフレーム毎に信頼できる **3** 次元位置を作成し，これを利用した **2D-3D** 対応から移動カメラの位置姿勢を推定する．実験により，本手法ではシーンの **3** 次元形状が未知の状態でも実画像列を利用し，カメラ位置姿勢が推定できることを確認した．

Keywords : **Camera Pose Estimation, Image Feature, Real Image**

1　はじめに

近年盛んに研究が行われ，同時に普及も進んでいる Augmented Reality(AR) においては，カメラの位置姿勢推定が重要な役割を占めている．AR においては，カメラやその使用者の実空間に対する位置･姿勢を把握することで，撮影した画像や使用者の視界に，3DCG などの新たな情報の重畳が可能になる．撮影した画像を基にカメラの位置姿勢を計算するためには，画像から得られる 2 次元位置と実空間の 3 次元位置の間での，正しい対応付けが求められる．このような 2D-3D 対応を作る手法として，事前にシーンのテクスチャ付き 3 次元形状モデルを利用する手法がある．この手法では，モデルから得られる局所特徴量に基づく特徴点データベースを利用し，移動カメラの位置姿勢推定を行っていた．ただし，このデータベースはモデルを多視点に投影した画像に基づいているため，実際の物体で起こる表面の見えの変化などの点で，実画像との間に差異を生じる．また，特徴点の 3 次元位置取得に用いるモデル形状の正しさの影響を受けるため，適用できるシーンに制約があった．

そこで本稿では，移動カメラの位置姿勢を推定するため，撮影対象となるシーンを多視点から撮影した実画像列を利用する手法を提案する．実画像から得られる特徴点・特徴量に基づいて位置の対応付けを行うため，対応付けの精度は従来手法より向上する．またシーン全体のテクスチャ・形状を含めた 3 次元復元は行わず，実画像同士の 2D-2D 点対応を起点としてカメラの位置姿勢を推定する．複数の 2D-2D 点対応から信頼できる 3 次元位置のみを計算し，これをカメラ位置姿勢の計算に利用する．対応付けされた 2D-2D 点対応は計算されたカメラ位置姿勢の精度評価にも活用し，最も良いカメラ位置姿勢の選択を行う．これにより，シーンの視点による見え方の変化や，形状の複雑さに起因する 3 次元復元の難しさの影響を受けない，カメラ位置姿勢の推定が可能となる．

2　関連研究

移動カメラの位置姿勢推定には，空間中の 3 次元位置と，それが画像に投影される 2 次元位置の対応が必要である．カメラ位置姿勢推定のためだけの作為的なマーカーを利用せず，マーカレスの手法でにこれを求めるには，視点が変化しても安定して検出できる位置をシーン中から取り出したい．シーンで点対応を安定して取得するために局所特徴量が多く用いられ，その例として SIFT[1] や SURF[2] などが挙げられる．しかし，これらは画像に射影変化が生じたとき，つまり撮影視点が大きく動いたときに特徴量の変化が大きい問題があり，このとき正しい点対応が安定して得られない．撮影視点の大きな変化に対応するため，Lepetit らの手法 [3] では姿勢推定の対象となる画像を射影変換した画像群をランダムに複数作成する．多くの画像で頻繁に特徴点検出される点を集め，それに対して入力画像の特徴点がどの点に対応するかを，周辺のパッチと決定木を用いて定めることで，姿勢変化に対応した正しい点の対応付けを行っている．また吉田らの手法 [4] では，同じく対象の平面画像を射影変換した画像群から局所特徴量を取り出し，これを利用する．彼

*1 慶應義塾大学
*1 Keio University

らの手法では，この画像群それぞれで特徴点を検出した上で，元の画像で同一の位置に対応する特徴点の特徴量群を保持している．これをデータベース化し，移動カメラで得る画像との間で特徴点を対応付けることで，視点が変化しても正しい点対応が頑健に得られ，カメラ位置姿勢の計算を可能にした．Tachasongthamら [5]，篠塚ら [6] はこの特徴点・特徴量保持の手法を，3 次元物体に拡張した．彼らはシーンの 3 次元形状モデルを作製し，ビューア上でその多視点からの投影画像を取得する Viewpoint Generative Learning (VGL) を行っている．取得した画像は擬似的にシーンを多視点から見たものとして扱い，これから局所特徴点を検出し，3 次元位置と画像特徴量を対応付けたデータベースを構築している．

しかし，このようなシーンのモデルを投影した画像は，同じ視点からの実画像とは異なる見えとなることが多い．テクスチャ付き 3 次元形状モデルの作製は，Structure from Motion (SfM) による手法があるものの [7][8][9]，正しく復元されるのは多視点で見える箇所に限られ，それ以外の箇所の形状にはゆがみが生じる場合がある．また，モデルのテクスチャは多視点での見えをマージしたものであるため，光の反射などによって見えが変化する場合には実際の見えとの間に差異が生じる．したがって，移動カメラで撮影された実画像由来の局所特徴量との間では，正しい対応付けが足りずに位置姿勢推定が失敗する場合がある．これを解決するためには，モデル形状の手作業での修正や，シーンの物体に対する材質・反射特性の設定，View-Dependent Texture Mapping[10] の利用などが必要となる．我々は，SfM で作製したシーンのモデルから特徴点の 3 次元位置のみを取得し，対応する画像特徴量はその入力である実画像列由来とする手法を提案した [11]．実画像を用いたために特徴量に基づく正対応数は多く得られるが，3 次元位置はモデルの形状に基づくためにその正確さの影響を受ける問題があった．

このように，シーン中の特徴点 3 次元位置や特徴量の取得にテクスチャ付き 3 次元形状モデルを利用すると，その復元精度が結果に影響することを考慮し，本稿ではモデルを利用しないカメラ位置姿勢推定を提案する．我々はシーンの環境を示すものとしてモデルに代わり，固定多視点のカメラで撮影した実画像列を用いる．この場合，移動カメラで撮影されるのは実画像であるから，局所特徴量に基づく点の対応付けは実画像同士でなされる．そのため，従来手法で扱っていたモデル投影画像由来の局所特徴量を利用した場合に比べ，より正しい点対応が得られる．

3 提案手法

提案手法全体の流れを図 1 に示す．本手法では，入力として移動カメラで撮影された画像 img^{IN} と，予めシーンを多視点から撮影した N 枚の実画像列を入力とする．これを実画像シーケンス $Img^{\mathrm{SQ}} = \{img^{\mathrm{SQ}_i} \mid 1 \leq i \leq N\}$ と以降呼ぶ．事前処理として Img^{SQ} について，それらを撮影したカメラの内部・外部パラメータ，及び各画像の特徴点群 $Kp^{\mathrm{SQ}_i} = \{kp_m^{\mathrm{SQ}_i} \mid 1 \leq m \leq M^{\mathrm{SQ}_i}, M^{\mathrm{SQ}_i}$ は各々の特徴点数 $\}$ と，その特徴量を取得する．

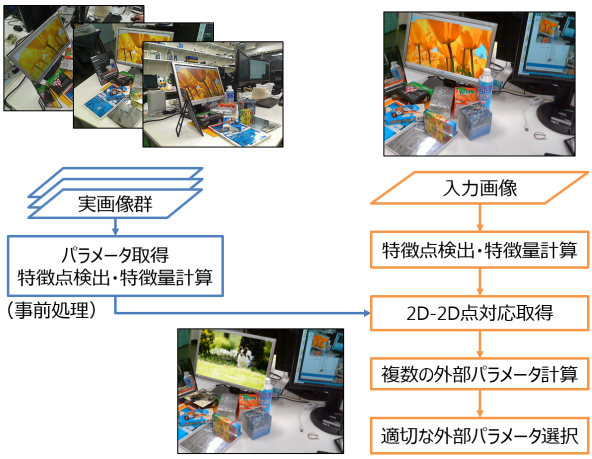

図 1　提案手法の流れ

3.1 正しい 2D-2D 点対応の取得

本手法では，まずはじめに img^{IN} について特徴点検出・特徴量計算を行う．このとき得られた特徴点が M^{IN} 個のとき，その点群を $Kp^{\mathrm{IN}} = \{kp_p^{\mathrm{IN}} \mid 1 \leq p \leq M^{\mathrm{IN}}\}$ とする．次に Kp^{IN} と $Kp^{\mathrm{SQ}_i}(1 \leq i \leq N)$ の間で，画像特徴量に基づく特徴点の対応付けを行う．この処理は特徴量のみを手掛かりとした対応付けであるため，幾何的には誤っている点対応も含まれている．そのため，最初に得られた 2D-2D 点対応から正しい点対応のみを取り出すため，RANSAC アルゴリズム [12] を利用した基礎行列の計算を行う．この結果，img^{IN} と img^{SQ_i} の間で正しい 2D-2D 点対応のみを得ることができる．なお，ここで計算される基礎行列は今後用いない．また，この処理を経てもなお誤った点対応が排除できない可能性があるが，これは 3.3 節で述べる処理で適切に扱われる．

3.2 複数の移動カメラ外部パラメータの計算

2D-2D 点対応において，img^{IN} の特徴点 $kp_p^{\mathrm{IN}}(1 \leq p \leq M^{\mathrm{IN}})$ と，実画像シーケンス img^{SQ_i} の特徴点 $kp_s^{\mathrm{SQ}_i}(1 \leq s \leq M^{\mathrm{SQ}_i})$ が，かつ kp_p^{IN} と img^{SQ_j} の特徴点 $kp_t^{\mathrm{SQ}_j}(1 \leq t \leq M^{SQ_j})$ が対応付けられているとする．この場合，$kp_s^{\mathrm{SQ}_i}$ と $kp_t^{\mathrm{SQ}_j}$ は同一の 3 次元位置を指すといえる．$img^{\mathrm{SQ}_i}, img^{\mathrm{SQ}_j}$ を撮影したカメラの

内部・外部パラメータは既知のため，三角測量によってこの点の 3 次元位置を計算できる．img^{SQ_i}, img^{SQ_j} の特徴点のうち，このように img^{IN} との 2D-2D 点対応を介して同一の 3 次元位置を指すことが分かる組み合わせが n 組あった場合，img^{IN} の特徴点の 2 次元位置と，計算された 3 次元位置の 2D-3D 点対応が n 組得られる．したがって，この点対応から Perspective-n-Point 問題を解き，移動カメラの外部パラメータ $\mathbf{Rt}_{ij}^{IN}\{1 \leq i, j \leq N, i \neq j\}$ を求める．ただし $n < 6$ の場合は Perspective-n-Point 問題を解けないため，$\mathbf{Rt}_{ij}^{IN}$ は求めない．

3.3 最も適切な移動カメラ外部パラメータの選択

3.2 節で求めた $\mathbf{Rt}_{ij}^{IN}$ について，i, j の組み合わせは ${}_N\mathrm{C}_2$ 通りある．したがって，移動カメラの外部パラメータは最大 ${}_N\mathrm{C}_2$ 個得られる．しかし，本手法で 2D-2D 点対応の取得のために初めに利用する画像特徴量は大きな視点変化に弱く，RANSAC アルゴリズムで求めた基礎行列と 2D-2D 点対応は，全てが正しいとは限らない．たとえば，大きく視点が異なる 2 視点間で画像特徴量に基づいて基礎行列を求めた場合，そもそも正対応が少なすぎて正しい基礎行列が得られず，正しい点対応と見なされる結果にも誤りが多く含まれることが多い．したがって，その結果を利用して求めた位置姿勢行列の精度にもばらつきが生じるので，最大 ${}_N\mathrm{C}_2$ 個の $\mathbf{Rt}_{ij}^{IN}$ のうち，最も精度が高いものを選択する必要がある．そのため最後の処理として，どの $\mathbf{Rt}_{ij}^{IN}$ が移動カメラの外部パラメータとして最も適切であるかを，$\mathbf{Rt}_{ij}^{IN}$ と実画像シーケンスの各外部パラメータ $\mathbf{Rt}_{k}^{SQ}\{k = 1, 2, \cdots N\}$ から求まる基礎行列を利用して選択する．

図 2 は，選択に利用する値を求める方法を示している．Rt_{ij}^{IN} を移動カメラの外部パラメータと仮定する場合，移動カメラのカメラ座標系を基準とする img^{SQ_k} を撮影したカメラの外部パラメータ $\mathbf{Rt}_{\mathrm{IN}-\mathrm{SQ}_k}$ は，以下の式で求められる．

$$\mathbf{Rt}_{\mathrm{IN}-\mathrm{SQ}_k} = \mathbf{Rt}^{\mathrm{SQ}_k}\mathbf{Rt}_{ij}^{\mathrm{IN}\,-1} \tag{1}$$

これに加えて img^{IN} と img^{SQ_k} を撮影したカメラの内部パラメータも利用し，2 つの画像座標間の基礎行列 $\mathbf{F}_{\mathrm{IN}-\mathrm{SQ}_k}$ が求められる．次に 3.1 節で求めた img^{IN}, img^{SQ_k} 間の 2D-2D 点対応のそれぞれについて $\mathbf{F}_{\mathrm{IN}-\mathrm{SQ}_k}$ を用い，img^{SQ_k} に img^{IN} の点に対応したエピポーラ線を引く．このとき img^{SQ_k} の画像座標上で，エピポーラ線と点の距離がしきい値以内である点対応の数をインライアとしてカウントし，その数を c_k とする．これは 2D-2D 点対応と $\mathbf{Rt}_{\mathrm{IN}-\mathrm{SQ}_k}$ の双方が正確であった場合，特徴点がエピポーラ線上に存在することを利用している．2D-2D 点対応は 3.1 節で RANSAC アルゴリズムを通して得ているために正しい点対応が多く含まれ，節での外部パラメータ計算に用いられなかった点対応もこの処理で活用できる．また，3.1 節で実際には誤っているにも関わらず正しい点対応と判定されたものは，この処理ではアウトライアとして扱われる可能性が高い．$k = 1, 2, \cdots N$ について $\mathbf{F}_{\mathrm{IN}-\mathrm{SQ}_k}$ の計算と，これを利用した正しい点対応数のカウントを行い，$\mathbf{Rt}_{ij}^{IN}$ の正確さ $ac_{ij} = \sum_{k=0}^{N} c_k$ を求める．以上の処理を最大 ${}_N\mathrm{C}_2$ 個の $\mathbf{Rt}_{ij}^{IN}$ に対して行い，ac_{ij} が最大となる $\mathbf{Rt}_{ij}^{IN}$ を，移動カメラの外部パラメータとする．

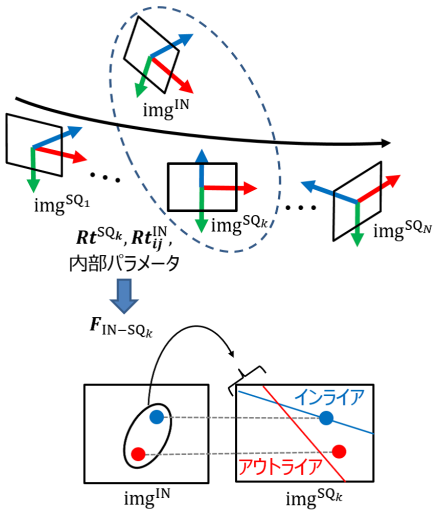

図 2　外部パラメータの精度評価のための
インライアの判定

4 実験

提案手法を利用し，カメラ位置姿勢推定の評価実験を行った．図 3 のようなシーンを用意し，これを撮影する移動カメラの位置姿勢を推定した上で，撮影された入力画像の任意の位置にテクスチャを重畳する．入力画像は移動カメラで撮影した 60 フレームとし，実画像シーケンスに含まれる画像の枚数 $N = 12$ とした．画像の解像度は，実画像シーケンス，入力画像共に 640×480 とした．比較のため，従来手法として実画像シーケンスを入力として作製した，図 4 のようなテクスチャ付き 3 次元形状モデルを基にして，VGL[5] によって画像特徴点データベースを作成した．これを用いて同様に，移動カメラ位置姿勢推定，テクスチャ重畳を行った．いずれの手法においても，画像特徴点検出・特徴量記述には SURF[2] を用いた．

また，重畳を想定するシーン中の位置（2 つのテクスチャのコーナー）を利用し，各手法の再投影誤差を評価した．再投影誤差は，最大 8 個のコーナーのうち画像中でその位置が分かる C 点を利用して計算する．入力画像中でコーナーが投影されるべき 2 次元位置を

$p_k(k = 1, 2, \cdots C)$ とする．これらの点は図 3 に示した 8 点とし，その 3 次元位置は実画像シーケンスから予め求めておく．各手法で推定されたカメラ位置姿勢によってこれらを入力画像に投影し，その 2 次元位置を $q_k(k = 1, 2, \cdots C)$ とする．i フレームの再投影誤差 err_i は以下の式 2 で表される．

$$err_i = \frac{1}{C}\sum_{k=1}^{C} \|p_k - q_k\| \tag{2}$$

図 3　実験シーン

図 4　シーンの 3 次元モデル

実験の結果，テクスチャを重畳した出力結果の一部を図 5 に示した．また，図 6 は 60 フレームの再投影誤差 err_i について，提案手法と従来手法を比較したものである．このように，多くの視点で提案手法は従来手法より重畳結果にばたつきが少なく，正しくカメラ位置姿勢推定がなされていることが分かる．

これは，従来手法ではデータベースに格納する画像特徴量とその 3 次元位置双方を，シーンの 3 次元形状モデルから得ていることが原因であると考えられる．モデルは視点による見え方の変化を考慮しないテクスチャを利用し，加えて細部では形状のゆがみが生じている．そのため，モデルの投影画像は実画像である移動カメラからの入力画像との間で見えに違いが生じ，特徴量に基づく正しい点対応が得にくい．また，データベース中で特徴量に対応した 3 次元位置はモデルの形状から取得しているため，形状が細部でゆがんでいることは，その周囲で取得される特徴点の 3 次元位置の不正確さの原因となる．そのため，提案手法より大きな再投影誤差が生じていると言える．

その一方で，提案手法では実画像列から特徴点・特徴量を取得するため，入力画像との間で正しい 2D-2D 点対応が得やすい．また，カメラ位置姿勢推定に必要なシーン中の 3 次元位置はその実画像同士の点対応から取得するため，従来手法より正しいものが得られていると言える．それに加えて図 4 に見られるように，復元した形状モデルは，シーンの一部のみしか 3 次元復元されていないため，復元されていない背景の部分では 3D-2D 点対応が得られない．提案手法ではその部分でも点対応が得られ，カメラ位置姿勢の推定に利用できるため，従来手法よりも空間的に広い範囲の活

図 5　移動カメラの位置姿勢推定結果
(左列：提案手法 右列：従来手法)

用が可能である．3.2 節で計算される，最終的に選択されるカメラ位置姿勢に利用されない正しい 2D-2D 点対応も多数存在するが，それらは 3.3 節で述べた，正しいカメラ位置姿勢の選択でインライアとしての役割を果たし，有効利用されている．

今後の課題としては，これらのインライアを最終的なカメラ位置姿勢の計算に利用する手法の検討がある．インライア全てを対象とし，エピポーラ線と対応する点の距離（誤差）の最小化をすることで，多くの点対応を移動カメラの位置姿勢の計算に利用できる．また，提案手法では入力画像と実画像列に含まれる N 枚 全てとの間で 2D-2D 点対応を求めているが，これらの中には視点が大きく離れた画像の組み合わせも含まれ，そこから得られる点対応には正しいものが少ない．そのため，前フレームで推定されたカメラ位置姿勢から大きく外れた実画像列中の画像を対応付けから除外し，

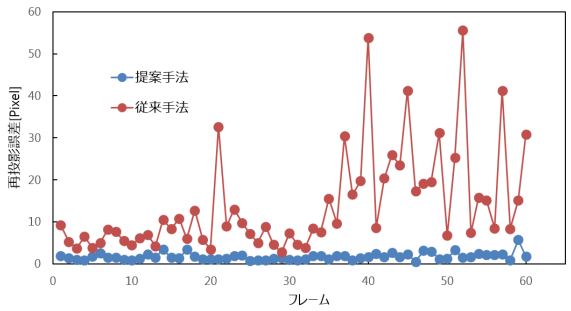

図 6　再投影誤差の比較

誤った点対応を減らしてより安定したカメラ位置姿勢が推定できる可能性がある．

5　まとめ

本稿では，AR に必要なシーンに対する移動カメラの位置姿勢推定を，シーンの 3 次元復元を行わずに実画像列を手掛かりに行う手法を提案した．提案手法では実画像同士から得られる点対応を利用したため，本来の見えや，形状の正確さが失われたテクスチャ付き 3 次元形状モデルを用いた従来手法よりも正確なカメラの位置姿勢推定を行うことができた．また，モデルでは正確な復元がなされない箇所で得られる点対応も，カメラの姿勢推定に活用した．

謝辞

本研究の一部は，科学研究費　基盤研究 (S)24220004 の補助により行われた．

参考文献

[1] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.
[2] H. Bay, T. Tuytelaars and L. V. Gool, “SURF: Speeded Up Robust Features,” in *Proc. 9th Europian Conference on Computer Vision*, May. 2006, pp. 404-417.
[3] V. Lepetit and P. Fua, “Keypoint Recognition Using Randomized Trees,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1465-1479, 2006.
[4] T. Yoshida, H. Saito, M. Shimizu, and A. Taguchi, “Stable Keypoint Recognition using Viewpoint Generative Learning’,” in *Proc. 8th International Conference on Computer Vision Theory and Applications*, Feb. 2013, pp. 310-315.
[5] D. Thachasongtham, T. Yoshida, F. de Sorbier and H. Saito, “3D Object Pose Estimation Using Viewpoint Generative Learning,” *Image Analysis*, vol. 7944, pp. 512-521, 2013.
[6] Y. Shinozuka, F. de Sorbier and H. Saito “Specular 3D object tracking by view generative Learning,” in *Proc. Irish Machine Vision and Image Processing Conference*, Aug. 2014, pp. 9-14.
[7] G. Klein and D. Murray, “Parallel tracking and mapping for small AR workspaces,” in *Proc. 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, Nov. 2007, pp. 225-234.
[8] C. Forster, M. Pizzoli, and D. Scaramuzza, “SVO: Fast semi-direct monocular visual odometry,” in *Proc. 2014 IEEE International Conference on Robotics and Automation*, Jun. 2004, pp. 15-22.
[9] J. Engel, T. Schöps and D. Cremers, “LSD-SLAM: Large-scale direct monocular SLAM,” *Computer Vision-ECCV 2014*, pp. 834-849, 2014.
[10] P. Debevec, C. Taylor, and J. Malik, “Modelling and Rendering Architecture from Photographs: A hybrid geometry- and image-based approach,” in*Proc. 23rd annual conference on Computer graphics and interactive techniques*, Aug. 1996, pp.11-20.
[11] K. Obata and H. Saito, “Camera Pose Estimation Based on Keypoints Matching with Pre-Captured Set of Real Images,” in *Proc. The Korea-Japan joint workshop on Frontiers of Computer Vision*, Feb. 2016, pp. 76-80.
[12] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381-395, 1981.