Vol.2015-CVIM-197 No.19
2015/5/18

実画像列を利用したカメラ位置姿勢推定のための安定特徴点データベースの作成

小畑 圭[1,a)]　斎藤 英雄[1]

概要：本稿では，カメラの位置姿勢推定を行いたいシーンにおいて，事前に対象シーンを RGB-D カメラにより様々な位置方向から撮影しておき，カメラの位置姿勢推定を実施するときに利用可能なデータベースを作成する手法を提案する．位置姿勢推定用のデータベースを事前に準備する手法として，これまでに筆者らのグループでは，シーンの 3 次元形状モデルを用いて視点変化に頑健な安定特徴点のデータベースを作成する方法を提案した．しかし，対象シーンの形状が複雑な場合など，シーンによっては形状モデルの入手・作成が難しいことがあり，このような場合には適用が困難であった．そこで本稿で提案する手法では，RGB-D カメラで直接対象シーンを多方向から連続で撮影した RGB 画像列・距離画像列を入力として，各視点の両画像から得られる特徴点の 3 次元位置情報をもとに視点の位置姿勢を計算し，安定特徴点の特徴量に基づくデータベースを作成する．そして，このデータベースが持つ特徴量と RGB カメラの入力画像から得られる特徴量に基づく特徴点マッチングによる 2 次元 3 次元対応から，RGB カメラの位置姿勢推定を行う．提案手法の有効性を確認するための実験により，本手法を用いることで 3 次元形状モデルを作成しなくても，従来手法と同等の精度でのカメラ位置姿勢推定が可能であることを確認した．

キーワード：カメラ位置姿勢推定，画像特徴量，3 次元形状

1. はじめに

近年，Augmented Reality(AR) の研究が進み，同時に普及も拡大している．視覚における AR では空間の様子をカメラを通して画像として認識し，画像上で実空間の情報を拡張する形で新たな情報を重畳する．このためには，カメラの実空間に対する位置姿勢を把握する必要がある．カメラの位置姿勢推定はシーン（実空間）の 3 次元情報と，シーンを撮影した画像の 2 次元情報の対応付けによってなされる．

本研究では，対象シーンについて視点変化に頑健な特徴量記述を持つ特徴点のデータベースを事前に作成し，カメラの位置姿勢推定を行う手法に着目した．従来手法ではシーンのテクスチャ付き 3 次元形状モデルを利用し，視点変化に頑健な特徴量を持つ安定特徴点 (Stable Keypoint) のデータベースを作成する．しかし正確な形状モデルは容易に手に入るものではなく，対象によっては入手・作成が難しい場合もある．

我々はシーンの形状モデルがない状況下で，RGB-D カメラで直接対象シーンを撮影し，Stable Keypoint のデータベースを作成する手法を提案する．提案手法では RGB-D カメラで直接，対象シーンを多方向から連続で撮影した RGB 画像列・距離画像列を入力とする．各視点の両画像から得られる特徴点の 3 次元位置情報をもとに，RGB-D カメラ視点の位置姿勢を計算し，Stable Keypoint 特徴量に基づくデータベースを作成する．実験により，提案手法を用いることで 3 次元形状モデルを作成せずとも，従来手法と同等の精度でのカメラ位置姿勢推定が可能であることを確認した．

2. 関連研究

マーカを用いないカメラの位置姿勢推定には，対象シーンの事前学習を行わない手法と，対象シーンの事前学習を行い，画像の自然特徴に基づくデータベースを保持しておく手法がある．事前学習を行わない手法の代表的なものとして，Klein らによる PTAM[1] がある．PTAM は画像からコーナー特徴を検出し，撮影で得られる連続した画像で安定して検出される特徴点をもとにシーンの 3 次元座標を決定する．PTAM では座標系決定に利用する 3 次元空間での 1 点は，その特徴が大きく変化しないことが前提である．そのため特徴の変化が大きい，急な位置姿勢変化には弱い．

[1] 慶應義塾大学
[a)] obata@hvrl.ics.keio.ac.jp

大きな位置姿勢変化に対応するには，様々な視点での対象シーンの見え方に対応できるデータベースを事前に作成する必要がある．Lepetit らの手法 [2] では，学習対象の画像をアフィン変換することで，視点変化時の見え方を考慮した画像をランダムに複数作成する．これらで頻繁に特徴点検出される点についてランダムにサブセットに分け，各々でどの特徴点であるかを定める決定木を作成する．入力画像の特徴点周辺のパッチを各決定木に通してどの特徴点らしいかを判定し，最も判定された数が多い特徴点とのマッチングとする．これにより，視点変化に対しても頑健なマッチングが可能となっている．

また Thachasongtham らの手法 [3] は，シーンの 3 次元形状モデルを入力として，その周囲に擬似的に視点を生成して得られた画像から，シーンの学習を行う．このとき多くの視点から同一の特徴点として検出される特徴点 (Keypoint) を，視点変化に頑健に検出される Stable Keypoint とする．そして，Stable Keypoint の 3 次元位置と，特徴点検出した全視点におけるその点の特徴量に基づく特徴量を，データベースに保存して位置姿勢推定に利用する．

Lowe らの手法 [4] では，対象を多方向から撮影した RGB 画像を入力とする．各画像の特徴点マッチングとエピポーラ拘束から，入力画像を撮影した視点間の相対位置を求めている．これより得られる 3 次元位置が求まる特徴点の集合をデータベースとして保存し，推定を行う．

本研究では，[3] で用いられた，Stable Keypoint を保持するデータベースについて，シーンの 3 次元形状モデルを必要とせずに作成する手法を提案する．

3. 3 次元形状モデルを利用したデータベースの作成

画像の局所特徴量として多く用いられる SIFT[5], SURF[6] は，視点変化に対して特徴量が大きく変化する性質がある．したがって，特徴点マッチングに用いる 3 次元点の特徴量としては不適切である．このため，視点変化に頑健に検出される Stable Keypoint について，視点ごとで異なる特徴量に対応できる特徴量を保持しておき，特徴点マッチングを行う必要がある．Yoshida らの手法 [7] では，Stable Keypoint の特徴量を，様々な視点の画像で得られた画像特徴量を基に記述する．Thachasongtham らの手法 [3] は Yoshida らの手法 [7] を 3 次元物体に拡張したものである．本章では，[3] の従来手法における Stable Keypoint 作成方法と，その特徴量に基づいたデータベース作成について述べる．また，作成したデータベースによるカメラ位置姿勢推定についても本章で説明する．

3.1 視点生成型学習による Stable Keypoint 作成

Stable Keypoint 特徴量は視点で変化する複数の画像特徴量を基に記述されるため，視点ごとのシーンの投影画像が必要である．従来手法では視点生成型学習（VGL）と呼ばれる手法で，Stable Keypoint の作成とその特徴量記述を行う．

VGL では図 1 に示すように，対象シーンの 3 次元モデルを用意し，周囲に擬似的に視点を生成する．視点の位置は，3 次元モデルが存在する 3 次元空間座標における角度 θ, ϕ, ω に依存する．特徴量記述子に SIFT を用いる場合，スケール変化・回転に不変な特徴量であるため，視点とその回転中心間の距離は問わず，回転角のうち ω も固定して問題ない．したがって θ, ω の値で定まる視点位置から見たシーンの様子を，モデルの投影画像として取得する．画像から特徴点検出・特徴量計算を行い，逆投影することで 3 次元位置と画像特徴量の関係が得られる．

このようにして得られた 3 次元点の画像特徴量から，Stable Keypoint を作成する．図 2 に示すように，各画像中で特徴点検出された 3 次元点を分布させる．その結果，図 2 の点 Q のように，多くの画像中で特徴点検出される 3 次元点が存在する．このような点は多視点での画像特徴量を持ち，検出される視点が多いほど視点変化に頑健な特徴点と言える．また，3 次元位置の距離がしきい値 Th_{kpt} 以下の特徴点は，同じ特徴点であると見なす．各視点の画像中で特徴点検出された全ての 3 次元点のうち，検出回数が上位の 3 次元点を Stable Keypoint として扱う．

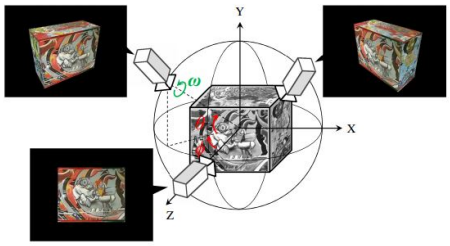


図 1　視点生成 [3]

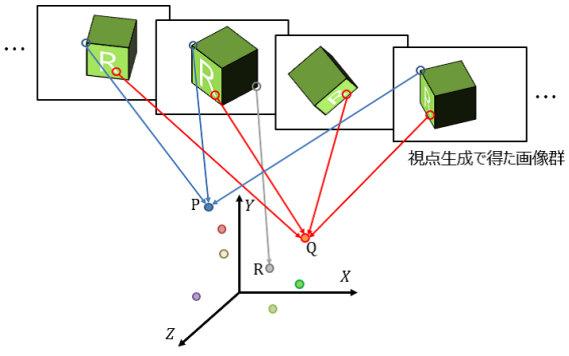


図 2　Stable Keypoint の作成過程

3.2 シーンのデータベース作成

Stable Keypoint の 3 次元座標と画像特徴量群から，トラッキング対象シーンのデータベースを作成する．カメラ位置姿勢推定時の探索効率化のために，Stable Keypoint 特徴量にはその点が持つ全ての画像特徴量を用いず，特徴

量を適当な数にクラスタリングする．[3] では，画像特徴量記述が N 次元であった場合に，その点が持つ画像特徴量群を N 次元空間で K-means クラスタリングし，K 個の特徴量をまとめて Stable Keypoint 特徴量として扱う．したがってクラスタリング数を K とするとき，1 つの Stable Keypoint はその 3 次元座標と K 個の特徴量記述を持つ．全ての Stable Keypoint についてこの処理を行うことで，特徴量と 3 次元座標の対応関係からなるデータベースが作成される．

3.3 データベースを用いたカメラ位置姿勢推定

カメラの位置姿勢推定は図 3 に示す流れで行う．入力は，シーンを RGB カメラで撮影して得られる画像 Img_{input} とする．データベース作成に用いたのと同じ画像特徴量記述子で Img_{input} から特徴点を検出し，特徴量を計算する．得られた全特徴点について，データベース中の全特徴量記述に対するマッチングを行う．このときマッチング高速化のため，あらかじめデータベースの特徴量についての探索木を作成しておく．データベース作成時に K-means クラスタリングを行った場合，探索木作成に Fast Library for Approximate Nearest Neighbors(FLANN) アルゴリズム [8] を利用する．

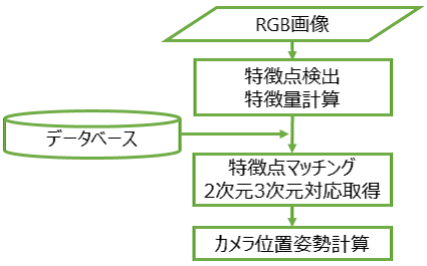


図 3 カメラ位置姿勢推定の流れ

マッチングは特徴量記述同士のユークリッド距離が最短のもので行うが，2 番目に近い特徴量記述を考慮する．データベース中の特徴量記述で最近傍との距離 d_1 と，2 番目との距離 d_2 を比べたとき，d_1 が d_2 に比べて小さくない場合は最近傍特徴量とのマッチングが誤りである可能性が高い．そこでしきい値 Th_{desc} を設定し，$d_1/d_2 < Th_{\mathrm{desc}}$ を満たす場合のみ，正しいマッチングとして扱う．なお，画像間の特徴点マッチングの場合と同様に，$Th_{\mathrm{desc}} = 0.6$ と定める．データベース中の特徴量記述は，3.2 で述べたように 3 次元座標と対応付けられている．したがって，Img_{input} で検出された特徴点の 2 次元画像座標と，データベース中の特徴点の 3 次元座標の対応が複数得られる．

最後に，これらの 2 次元 3 次元対応をもとに RGB カメラの位置姿勢を示す行列の計算を行う．カメラの位置姿勢は，シーンが持つ 3 次元世界座標系から 3 次元カメラ座標系への変換行列 $\boldsymbol{Rt}$ によって表される．カメラの内部パラメータを $\boldsymbol{A}$，画像座標系での特徴点の 2 次元座標を $\boldsymbol{m}$，マッチングする 3 次元座標を $\boldsymbol{X}_W$ とすると，式 (1) の関係が得られる．

$$\tilde{\boldsymbol{m}} \sim \boldsymbol{A}(\boldsymbol{I} \mid \boldsymbol{0})\boldsymbol{Rt}\tilde{\boldsymbol{X}}_W \tag{1}$$

式中の $\sim$ は定数倍の不定性を許すことを表すので，不定性を含む部分を 1 に正規化することで，$\boldsymbol{A}(\boldsymbol{I} \mid \boldsymbol{0})\boldsymbol{Rt}$ に含まれる未知数の数は 11 である．したがって最低 6 組の 2 次元 3 次元対応により，$\boldsymbol{Rt}$ の算出が可能である．得られた対応が 6 組を超える場合，計算に RANSAC[9] を利用する．これによって Th_{desc} で除去できなかった誤ったマッチングは外れ値として扱われ，より正しい推定を行える．

4. 実画像列を利用したデータベースの作成

3 章で述べた従来手法では，Stable Keypoint 作成に対象シーンの 3 次元形状モデルを必要とした．そのため，形状モデルがない状態では Stable Keypoint を作成出来ない．加えて，シーンの正確な形状モデルの作製には時間と労力を要する．直方体のような単純な物体を対象シーンとする場合は，各面のテクスチャを用意すればモデル作製は難しくない．しかし，物が乱雑に置かれた環境や複雑な形状を対象とするとき，形状モデルの作製は容易ではない．

したがって，3 次元形状モデル作製をすることなく Stable Keypoint の作成を行い，データベース作成を行えることが望ましい．本章では提案手法である，RGB-D カメラで対象シーンを撮影した実画像列を利用した Stable Keypoint とシーンのデータベース作成について述べる．提案手法の流れは図 4 のようになっている．

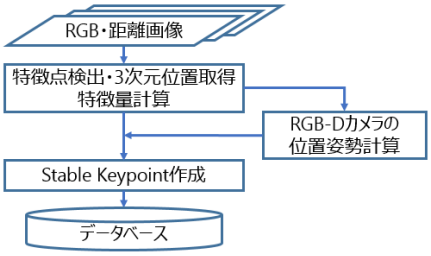


図 4 提案手法の流れ

4.1 RGB-D カメラの位置姿勢計算

入力として，対象シーンを RGB-D カメラで連続的に撮影した，全視点での RGB 画像・距離画像を用いる．各視点において，使用する画像特徴量記述子に基づく特徴点を RGB 画像から検出する．本手法では SIFT[5] を利用した．特徴点位置は 2 次元画像座標で示されるが，距離画像を利用してその 3 次元位置を取得する．今後，本稿では「特徴点の位置」は特徴点の 3 次元座標を示す．

撮影した RGB 画像群のうち，視点が異なる 2 つの画像について，特徴点マッチングを行う．選択する 2 つの画像は，撮影が時間的に連続しているように，RGB-D カメラの位置・姿勢共に大きな変化がない状況であるとする．こ

の場合，視点変化に弱い画像特徴量でも，マッチングには空間内で同じ位置を指すものが多数含まれる．マッチングの結果，次に示すような複数組の 3 次元の特徴点対応が得られる．

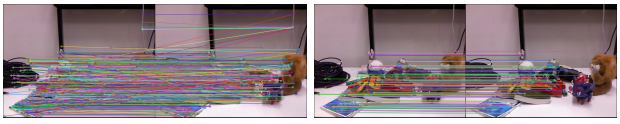
$$\left\{ \begin{array}{l} A_1(a_{1X}, a_{1Y}, a_{1Z}) \\ A_2(a_{2X}, a_{2Y}, a_{2Z}) \end{array} \right., \left\{ \begin{array}{l} B_1(b_{1X}, b_{1Y}, b_{1Z}) \\ B_2(b_{2X}, b_{2Y}, b_{2Z}) \end{array} \right., \cdots$$

この関係を用いて，式 (2) より 2 つのカメラ座標の変換行列 $\boldsymbol{Rt}$ を算出する．

$$\begin{bmatrix} a_{1X} & b_{1X} & \cdots \\ a_{1Y} & b_{1Y} & \cdots \\ a_{1Z} & b_{1Z} & \cdots \\ 1 & 1 & \cdots \end{bmatrix} = \boldsymbol{Rt} \begin{bmatrix} a_{2X} & b_{2X} & \cdots \\ a_{2Y} & b_{2Y} & \cdots \\ a_{2Z} & b_{2Z} & \cdots \\ 1 & 1 & \cdots \end{bmatrix} \quad (2)$$

ただしマッチングの結果には誤対応が存在するため，これを除去した計算を行う必要がある．RANSAC[9] を用いて誤対応を除去し，正しい変換行列を取得する．

まずマッチングで得られた 3 次元点対応から 4 組をランダムに選択し，仮の変換行列 $\boldsymbol{Rt}_{\mathrm{tmp}}$ を算出する．次に 4 組を除く全てのマッチングに $\boldsymbol{Rt}_{\mathrm{tmp}}$ を適用し，3 次元位置の誤差がしきい値 Th_{dist} 以内である組み合わせ数 $score$ を記録する．このような 4 組の選択から $score$ 記録までの処理を $loop$ 回行い，$score$ が最大であった $\boldsymbol{Rt}_{\mathrm{tmp}}$ を $\boldsymbol{Rt}_{\mathrm{Maxscore}}$ とする．最後に，$\boldsymbol{Rt}_{\mathrm{Maxscore}}$ を適用して 3 次元位置の誤差がしきい値以内である全てのマッチングを式 (2) に当てはめ，計算される変換行列をこの 2 視点間の $\boldsymbol{Rt}$ とする．図 5 は，2 視点の RGB 画像の特徴点マッチング結果を示したものである．(a) は全てのマッチングを，(b) は RANSAC を利用して抽出したマッチングのみを表示している．このように，$\boldsymbol{Rt}$ の計算は正しいマッチングのみを用いていることが分かる．



(a) (b)

図 5 マッチング精度の比較

以上の処理を全ての視点について行うことで，入力画像列を撮影した全視点のカメラ座標系について，2 視点間の位置姿勢の関係を示す行列が求まる．全ての $\boldsymbol{Rt}$ を隣接視点間で求めた場合，各視点について $\boldsymbol{Rt}$ を順に掛け合わせることで，入力画像列の端の画像を撮影した基準視点への変換行列が求まる．基準視点のカメラ座標系への変換行列により，各視点の全特徴点の位置を，統一した 3 次元座標系で表すことができる．

4.2 Stable Keypoint の作成とデータベース化

特徴点の位置・特徴量を利用して，Stable Keypoint を作成する．各視点について，4.1 節で求めた変換行列により，特徴点 kpt_{tmp} の位置を基準視点のカメラ座標系で表現する．既に存在するすべての Stable Keypoint 候補の中で，kpt の最近傍点と 2 番目に近い点を $skpt_{\mathrm{nearest}}$，$skpt_{\mathrm{second}}$ とする．

しきい値 Th_{kpt} 以下の距離に $skpt_{\mathrm{nearest}}$ が存在する場合，kpt_{tmp} は $skpt_{\mathrm{nearest}}$ の構成要素であると判定する．$skpt_{\mathrm{nearest}}$ と kpt_{tmp} の位置をそれぞれ P_{nearest}，P_{tmp} とし，$skpt_{\mathrm{nearest}}$ を構成する特徴点の数を $vote_{\mathrm{nearest}}$ とするとき，P_{nearest} を式 (3) のように更新する．

$$\overrightarrow{OP_{\mathrm{nearest}}} = \frac{vote_{\mathrm{nearest}}\overrightarrow{OP_{\mathrm{nearest}}} + \overrightarrow{OP_{\mathrm{tmp}}}}{vote_{\mathrm{nearest}} + 1} \quad (3)$$

このとき $skpt_{\mathrm{nearest}}$ の特徴量に kpt_{tmp} の画像特徴量を付加し，$vote_{\mathrm{nearest}}$ に 1 を加える．

$skpt_{\mathrm{nearest}}$ との距離が Th_{kpt} 以上であり，かつ $skpt_{\mathrm{second}}$ との距離が $2Th_{\mathrm{kpt}}$ 以上である場合，位置を kpt_{tmp} の 3 次元位置，特徴量を kpt_{tmp} の画像特徴量，$vote_{\mathrm{kpt}} = 1$ として新たな Stable Keypont 候補を作成する．

各視点の全特徴点についてこの処理を行うことで，RGB-D カメラで撮影した全視点での見え方を考慮した特徴量記述を持つ Stable Keypont 候補が作成される．このうち $vote$ の値が上位の候補を Stable Keypont として扱い，カメラ位置姿勢推定のデータベース作成に使用する．データベースの作成は，3.2 節に示した従来手法 [3] と同様の手順である．

5. 実験

本章では，提案手法と従来手法 [3] で作成したデータベースによる，カメラ位置姿勢推定の比較実験について述べる．位置姿勢推定は，両手法ともに 3.3 節の手法を用いた．

5.1 実験の概要

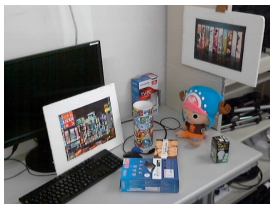


図 6 実験の対象シーン

図 7 3 次元形状モデル

対象のシーンは，図 6 に示すような環境とした．このシーンを RGB カメラで撮影した動画を入力とし，フレームごとにカメラ位置姿勢を推定した．データベース作成において，提案手法では RGB-D カメラとして Kinect v1 を使用し，従来手法では Autodesk 社が提供するアプリケーションである 123D Catch[10] を利用して作成した，図 7 のような 3 次元形状モデルを使用した．データベース作成で設定したパラメータの値は次の通りである．

- 提案手法・従来手法で共通
  - Th_{kpt}: 8.0mm

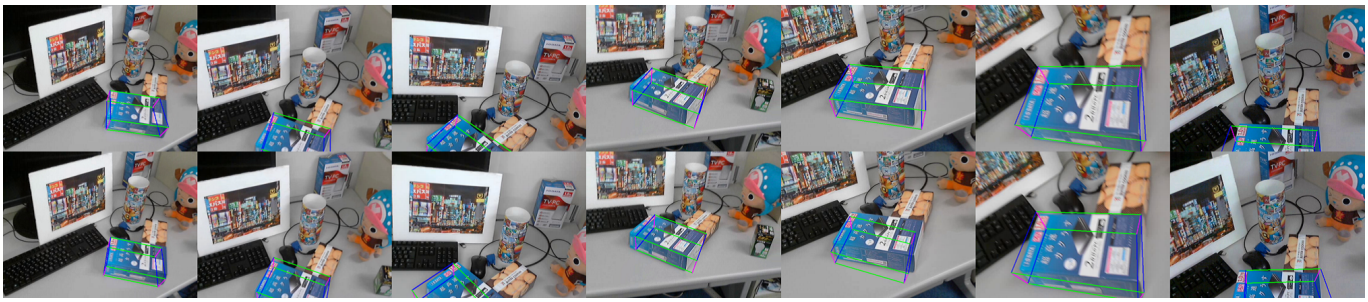


図 8　カメラ位置姿勢推定結果（上段：提案手法 下段：従来手法 [3]）

– データベース中の Stable Keypoint 数: 2000
– Stable Keypoint の画像特徴量クラスタリング数: 8

- 提案手法
  – Th_{dist}: 2.0mm
  – *loop*: 5000
- 従来手法
  – 視点生成におけるサンプリングの間隔: 5.0°

5.2　実験結果画像

図 8 は出力結果の一部であり，青い箱を囲む線を描画している．上段が提案手法，下段が従来手法で作成したデータベースを使用したものである．

5.3　考察

図 8 に示されるように，提案手法で作成したデータベースは従来手法と同等の精度でカメラ位置姿勢推定が実現できていることが分かる．

本手法では実画像をデータベース作成に用いたが，このことは視点により見えが変化するシーンに対する位置姿勢推定への活用が考えられる．たとえば光沢がある表面は視点により見えが大きく変化するが，3 次元形状モデルではその様子の再現が難しく，VGL における擬似視点から得た画像では実際の見えと異なる．したがってカメラ位置姿勢推定時に，実画像である入力画像と，データベース間の特徴点マッチング精度が低い．一方で提案手法では，データベース作成時に実際の見えを考慮した実画像列を入力とするので，位置姿勢推定時のマッチングは，実画像で記述された特徴量同士を基にしてなされる．このことからマッチング精度が向上し，見えの変化に対応した位置姿勢推定を行えることが期待される．

6.　まとめ

本研究では，対象シーンを RGB-D カメラで撮影することでカメラ位置姿勢推定用のデータベースを作成した．このとき Stable Keypoint 特徴量を利用することで，視点の変化に対する頑健さを確保した．さらに従来手法では Stable Keypoint 作成に必要とされていた，シーンの 3 次元形状モデル作製を介さずに Stable Keypoint の作成を行えた．

実験の結果，提案手法のデータベースは従来手法のデータベースとほぼ同程度の推定を行えることがわかった．したがって，従来より簡易な処理で Stable Keypoint 特徴量を用いたデータベース作成ができ，カメラ位置姿勢推定に利用できた．

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] 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.
[2] 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.
[3] 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.
[4] I. Skrypnyk and D. G. Lowe, “Scene Modelling, Recognition and Tracking with Invariant Image Features,” in *Proc. Third IEEE and ACM International Symposium on Mixed and Augmented Reality*, Nov. 2004, pp. 110-119.
[5] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.
[6] 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.
[7] 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.
[8] M. Muja and D. G. Lowe, “Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration,” in *Proc. International Conference on Computer Vision Theory and Applications*, Feb. 2009, pp. 331-340.
[9] 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.
[10] Autodesk 123D Catch ⟨http://www.123dapp.com/catch⟩ (2015/4/15)