Vol.2011-CVIM-176 No.1
2011/3/17

多視点3次元復元の研究動向

鳥 居 秋 彦[†1]　岡 谷 貴 之[†2]　延 原 章 平[†3]

多視点画像を使って物体やシーンの 3 次元形状およびカメラの姿勢を得る問題は，コンピュータビジョンの中心的課題として長い間研究されてきた．活発な研究活動を通じて得られた成果は，ロボット視覚から，CG や拡張現実 (AR) を始めとする映像メディア応用まで，様々な形で実社会にフィードバックされている．本稿は，多視点 3 次元復元の最近の研究動向をサーベイしたもので，大量未整列画像からの SfM，実時間 SfM/Visual SLAM，人物の全周囲 3 次元形状・運動復元の 3 つのトピックについて述べる．

Recent Research Trends in Multi-View Three-Dimensional Reconstruction

Akihiko Torii,[†1] Takayuki Okatani[†2]
and Shohei Nobuhara[†3]

The problem of obtaining the three-dimensional structure of an object or a scene and camera poses from multiple view images is one of the central issues of computer vision and has been studied for a long time. The results of intense research activity in the area have been applied to various real-world problems from robotic vision to image media application such as CG and augmented reality. This article surveys recent research trends in multi-view three-dimensional reconstruction, focusing on three topics, SfM for unordered image collections, realtime SfM/Visual SLAM, and complete three-dimensional recovery of human body shape and motion.

1. 大量未整列画像からの SfM

インターネット上に存在する膨大な量の画像[28)21)27)] を用いることで，世界そのものを 3 次元復元することも遠い未来の話ではない．最先端のコンピュータビジョン技術を集結することで，画像情報のみを用い，街という規模の復元が可能であることが，実証されている[2)]．インターネットを通して得られる画像は，様々な人々，時刻，季節，カメラで撮影されているため，画像のタイムスタンプ等から時間情報を得られたとしても，ビデオ映像のような時空間情報に対する仮定は成り立たない．本章では，そのような大量の未整列画像から 3 次元復元を行うための基本的な要素技術と最近の動向に関するサーベイを行う．

はじめに，本章で言う ”復元 (reconstruction) ”問題とは，複数の画像から，

- カメラの位置姿勢 (camera pose)
- ある 3 次元空間点がどのカメラから見えているかを表す対応関係 (correspondence)
- 撮影されたシーンに含まれる (疎な) 3 次元空間点群

を推定することを指し， Structure from Motion (SfM) と同義語と捉えて頂いても良い．Dense reconstruction[18)] や，surface reconstruction[9)] は，本サーベイには含まないことにする．

数十枚という小規模な未整列画像データセットに対する復元は，Schaffalitzky-Zisserman らが，ECCV2002 論文[62)] で成功しており，これから紹介する研究の多くはこの発展型として捉えられる．しかしながら，この論文で用いられている各要素技術は，必ずしも現在の標準的手法ではない．一方，近年の 3 次元復元研究で欠かせないソフトウェアの代表として，Snavely による Bundle[67)65)] がある．Bundler は，Phototourism[64)] や，Photosynth[44)] の根幹を担う SfM ソフトウェアであり，Bundler で用いられている要素技術は，他の様々な復元システムと共通部分が多い．したがって，Bundler で用いられている復元の流れを軸に， 3 次元復元の基本的なシステムを概略すると同時に，各要素技術の動向を紹介する．

未整列画像からの復元では，はじめにどの画像が同じ対象を撮影したものであるかを認識

†1 東京工業大学
Tokyo Institute of Technology
†2 東北大学
Tohoku University
†3 京都大学
Kyoto University
⋆1 鳥居が第 1 節，岡谷が第 2 節，延原が第 3 節をそれぞれ担当した．

© 2011 Information Processing Society of Japan

する必要がある．シンプルな方法として，各画像ペアごとにマッチングを行い，画像の接続関係を示す image connectitity graph を作成する．Image connectivity graph から初期復元を行うペアを選択し，カメラ位置姿勢と３次元空間点の復元を行う．復元済みの点と共有視野を持つ画像を追加，復元点の追加，バンドル調整を繰り返す．次節では，最初の課題にあたる correspondence problem について紹介する．

1.1 Correspondence Problem

何を特徴とするか？特徴的な特徴とはなにか？特徴量としてどういう記述をするべきか？という問題が長きに渡って研究されてきた[22)40)]．画像間で対応する２次元点を探索する問題は，対応づけ問題 (correspondence problem) と呼ばれており，同じシーンを撮影した画像間で，３次元物体の見かけ (appearance) は似ているという仮定することで，自動での対応付けが可能である．

ビデオ映像など連続フレーム間では，各画像間におけるカメラ運動が十分に小さく，見かけの変化が非常に小さいという仮定が成り立つ．その仮定のもとに，勾配画像から画像中の特徴的な情報 (コーナーなど) を選択し，それらの周辺画素が画像間で一致する条件を元に対応付けを行うアプローチが一般的であった[22)40)60)]．ただし，３画測量による復元精度は，カメラ間の並進運動量に依存するため，見かけの変化が少ないという仮定を満たしつつ，カメラの運動量を確保するような撮影条件は限定される．

未整列画像間で対応付けを行う場合，画像間のカメラ運動が小さいという仮定は適切ではない．さらに，様々な時刻，季節での撮影が想定されるため，見かけの変化が大きな画像間での対応付けが必要であり，通常，wide baseline stereo (WBS) matching[5)42)] が用いられる．WBS matching の基本的なアルゴリズムは，以下の通りである．

(1) 特徴検出と記述子 (特徴ベクトル) の生成 (feature detection and description)
(2) 特徴の対応付け (feature matching)
(3) Geometric verification

以降，２枚の画像間特徴の対応付けを”feature matching”，複数枚での画像間特徴の対応付け”tracking” と呼ぶことにする．以下，WBS matching の各ステップで用いられるアルゴリズムについて紹介する．

1.1.1 Feature Detection and Description

より様々な画像変換に対応可能な特徴量検出器，記述子 (feature detector, descriptor) が長きに渡って求められていた．そのような detector と descriptor アルゴリズムが確立されたことにより，wide baseline stereo (WBS) matching が可能になったといっても過言ではない．

Scale Invariant Feature Transform(SIFT)[39)] は，画像の回転，拡大縮小，輝度やコントラストの変化，を含む変換に対し頑健な feature detector, descriptor である．SIFT は，画像をダウンサンプルすることでスケールの異なる画像を用意し，それらに Gaussian filter をかけ，平滑化画像を生成する．さらに隣り合うスケール間で平滑化画像の差分を計算することで，スケール変化にたいして，不変な小領域 (blob) の中心や，コーナー等を検出する．その周辺領域の勾配方向のヒストグラム計算することで，輝度やコントラストの変化に対して頑健なを記述子を生成する．

SIFT の発展形として，box filter と integral image を利用することで，１０倍近く高速に，SIFT に近い精度でマッチングを行う Speeded Up Robust Features(SURF)[6)]，SIFT で検出した局所領域の勾配情報に対して主成分分析を適用し，より頑健かつ低次元な記述を実現する PCA-SIFT[29)]，より頑健な特徴記述子である GLOH[46)] など様々な手法が提案されている．SIFT 等に関する詳細は，藤吉のチュートリアル[81)] を参照されたい．

さらに，より広いクラスの画像変換：アフィン変換により共役に変化する特徴量 (affine covariant region) として，Maximal Stable Extremal Region(MSER)[42)]，Harris-Affine, Hessian-Affine[45)] などがある．一例として，MSER は，輝度値に対して段階的な閾値処理を施したときに安定な (停留する) 領域を検出する．そして，境界を抽出し記述すべき楕円領域 local affine frames(LAF)[55)] を設定し，descrete cosine transform(DCT) などを用いて記述子を生成する．Mikolajczyk らによる affine covariant region detector についてのサーベイ論文[47)] では，多くの実験において，detector として，MSER が照度・視点変化に対して最もロバストであると報告している．SIFT, SURF, MSER は，OpenCV ライブラリ[7)] に組み込まれており，またその他にも様々なプログラムが公開されている[77)76)]．

大量画像からの３次元復元を行うシステムの多くは SIFT を採用している[65)2)25)],[3)19)]．魚眼レンズカメラ画像や，Streetview などのパノラマ画像を用いる復元システムでは，画像間での特徴量変化が大きいため affine covariant regions も利用されている[25)26)]．

1.1.2 Feature Matching

SIFT を用いた場合，各特徴量は６４次元のベクトルとして表現される．２枚の画像間で検出された各特徴量の対応付けるには，最も類似しているベクトル探す，つまり，最近傍探査を行う．ここで扱う feature matching は，tentative matching，putative matching とも呼ばれる．

高速に最近傍探査を行うために， kd-tree を用いた approximate nearest neighbours

© 2011 Information Processing Society of Japan

(ANN)[4] や，最近では，fast library for approximate nearest neighbors(FLANN)[48] が用いられており，OpenCV[7] にも組み込まれている．尚，FLANN では，randomized kd-tree アルゴリズムと hierarchical k-means tree アルゴリズムが実装されている．さらに，アルゴリズムの種類，ランダム木の数，k-means 木における分枝の数や反復回数をパラメータ，探索時間，木構築にかかる時間，メモリ使用量からなるコスト関数を最適化することで，データセットの構造に応じて最適な探索アルゴリズムとそのパラメータも導出してくれる．

1.1.3 Geometric Verification

特徴ベクトルの対応付けにより画像間での特徴点対応を得られるが，特徴ベクトルの類似度のみに基づいて構成された対応点は，誤対応を含む可能性が高い．とはいえ，ひとたび対応点が決まれば，カメラの相対的な運動を推定することができる．そこで，仮定生成とその検証を繰り返し，最もコンセンサスの高いモデルとそのサポートを出力する Random Sample Consensus(RANSAC)[16] が用いることで，誤対応除去とカメラ運動の推定を同時に行うことが可能である．RANSAC を以下に簡潔にまとめる．

(1) 標本点 (対応点) をランダムに選択
(2) 幾何モデル (homography, fundamental matrix など) を計算
(3) その幾何モデルの検証 (サポート数の評価など)

以上の３ステップを一定数反復する．反復回数の設定については,[16] または,[23] を参照されたい．幾何モデルの計算法は，次節で紹介する．この RANSAC スキームに関して，その効率，頑健性の向上を図り様々な研究が行われてきた．

標本数が非常に多い場合，RANSAC における検証のステップの計算コストが高くなる．R-RANSAC[43] では，以下の工夫により，検証にかかるコストを大幅に削減する．

(1) 全標本からランダムに抽出した部分集合に対してのみ検証を行う
(2) ステップ１を通過した仮定のみ，残りの標本に対する検証を行う

さらに，Optimal Randomized RANSAC[13] では，順次標本の検証を行う際，sequential probability ratio test(SPRT) を用いることで，より効率よく ”悪い ”仮定を棄却している．

一般的な RANSAC では，反復ごとにある仮定に対し最後まで検証を行うという意味で，深さ優先アルゴリズムである．それに対し，Preemptive RANSAC[51] は幅優先型のアルゴリズムになっている．Preemptive RANSAC を概略すると以下の通りである．

(1) ５００個の仮定を生成，対応点を１００ずつのブロックに分割
(2) ステップ (4)，(5) を反復
(3) 全仮定に対し次のブロックの１００個の対応点を検証
(4) 累計サポートの少ない仮定を削除

個数，評価数は例である．例えば，ビデオ映像が入力であり，次の画像ペアを計算するまでの一定時間が確保されている場合，Preemptive RANSAC は非常に有効なアルゴリズムである．RANSAC スキームの中では，計算量を固定した中で最適なモデルを出力するというアイデアは，とても新鮮であった．

SIFT などを用いて生成された対応点では，当然その類似度 (descriptor 間のユークリッド距離など) が計算できる． RANSAC において，無作為にサンプリングをするのではなく，類似度の高いものから順にサンプリングを行うほうが多くの場合効率的である．そのようなサンプリングを取り入れた RANSAC が Progressive Sample Consensus(PROSAC)[12] である．PROSAC におけるサンプリングでは，反復終了条件を満たせず，反復回数上限に到達した場合，ランダムサンプリングと同じサンプリングパターンとなる，すなわち，サンプリングに偏りが生じないよう工夫がしてある．

さらに， Adaptive Real-Time Random Sample Consensus(ARRSAC)[59] は， Preemptive RANSAC，PROSAC，SPRT を融合した，現段階での最先端 RANSAC アルゴリズムであるといえる．

上記で紹介した研究以外にも，MLESAC[74] は，単純にサポートの数でモデルを評価するのではなく，尤度を評価することで，頑健性を増している．幾何学的な縮退 (degeneracy) が生じていないかを検証しつつ安定な解をもとめる degensac[14]，予め tentative match をグルーピングすることで頑健性を向上する GroupSAC[49]，予め周囲の対応点と類似度を測り discriminative な tentative match のみを用いる SCRAMSAC[61]，RANSAC でサポートを許容する閾値 (tolerance) を動的に更新する StaRSAC[10] など現在も様々アルゴリズムが提案されている．

1.2 Computing Camera Poses

RANSAC では，画像間の点対応が取れたと仮定して，カメラ間の相対的な運動を陰に求めている．本節では，対応点からカメラの位置関係を求めるアルゴリズムを紹介する．

画像から (up to similarity transformation での) 復元を行う際，焦点距離，歪み係数など，カメラの内部パラメータ情報が必要である[23]．カメラキャリブレーションの詳細については,[80] を参照されたい．インターネット上にアップロードされた画像を用いて復元を行う場合，ユーザー自身がカメラキャリブレーションを行うことは通常起こりえない．そこで，撮影時に記録された画像のメタデータ (EXIF など) に保存されている焦点距離等をを初期値として用い，後のカメラ位置姿勢，バンドル調整の段階で再推定を行うのが一般的で

© 2011 Information Processing Society of Japan

ある[65],[35]．EXIF データから焦点距離を計算する手順は，Snavely の博士論文付録を参照のこと[68]．

画像からの復元問題においては，camera poses も 3D points も未知であるが，各画像のある画素を通る光線が３次元空間で交わるという幾何学的拘束条件を用いることで，両者を同時に推定することが可能である．具体的には，空間の同一点が２つのカメラに投影されたときに成立するエピポーラ拘束 $\mathbf{x}^\top \mathtt{F}\mathbf{x} = 0$ から，十分な数の対応点を用いて fundamental 行列 $\mathtt{F}$(以下，F 行列) を計算することで，カメラの相対的な位置関係が推定できる．８点以上の対応点が得られる場合，F 行列は，8 点アルゴリズムを用いて計算される[36][23]．実用上は，画像ノイズ，解像度差などにより頑健な normalized 8-point algorithm[24] が用いられる．さらに，F 行列の rank が２であるという拘束を用いることで７組の対応点から F 行列を計算することが可能である[23]．

カメラの内部パラメータが既知であり，対応点がキャリブレート済みの場合，２組の対応点間に存在する幾何学的拘束は，カメラ間の相対的な回転と並進運動から構成される Essential 行列 (以下，E 行列) を用いて表される．ここで，エピポーラ拘束と，E 行列の２つの特異値が１であり，最後のひとつが０であるという拘束条件を用いることで，５組の対応点から E 行列が計算できる[30]．上述した拘束条件から E 行列を求めるには，１０次元多項式を解く必要があり，実用的なアルゴリズムが長らく存在しなかったが，Nister がグレブナ基底と action matrix を用いた効率的な解法，５点アルゴリズム[52] を提案した．それ以降，いくつか他の解法も提案されている[34][31]．

RANSAC において，５点アルゴリズムは，モデル推定に必要な対応点数が８点アルゴリズムに比べ少ないことから誤対応を含む確率が低く，効率よく解を発見しやすい．さらに，解の精度も良いことが[34] 報告されている．また，実用面において，点とカメラの配置によって生じる縮退 (degeneracy) が少ないことも重要な利点である[⋆1]．

５点アルゴリズム以外にも，カメラ運動と焦点距離を直接求める６点アルゴリズム[69]， ４組の対応点を用い３視点のカメラ位置姿勢を推定するアルゴリズム[53][58][33] など，各問題に合わせた解法がある．詳しくは[71] を参照されたい (各ソースコードへのリンク有)．

ここでカメラの位置姿勢が既知であれば， ３角測量 (triangulation) によって空間点位置を計算できる．一方，空間点位置が既知であれば，resectioning[23] によって，カメラの位置姿勢を計算できる．内部パラメータが既知の場合は P3P 問題[16] と呼ばれ，空間の３点とその投影点 (対応点) を用いてカメラ位置姿勢を求める．最近では，ジャイロや加速度センサから得られる重力方向を用いることで，空間の 2 点とその投影点からカメラ位置姿勢をクローズドフォームで求める解法[32] なども提案されている．内部パラメータが未知の場合は，DLT アルゴリズム[23] を用いることで，焦点距離，カメラ中心等を求めることも可能である．尚，Bundler[65] では， DLT で求めた焦点距離が EXIF 焦点距離に十分近い場合のみ，その結果を採用している．

以降，２枚の画像に対し，feature detection and description, feature matching, RANSAC と geometric verification まで一連の処理を画像ペアマッチング (image pair-wise matching) と呼ぶ．

1.3 バンドル調整

未整列画像からの復元においても，５点アルゴリズムや DLT アルゴリズムにより得られたカメラ位置姿勢，空間点位置を初期値として，それらの refiment を行うモジュールとして，バンドル調整が用いられる．ここで，復元された空間点が画像面上に投影された点を再投影点，その空間点に対応する画面上で検出された特徴点を観測点と呼ぶ．バンドル調整とは，最投影誤差 (観測点と再投影点間のユークリッド距離) の２乗和をコスト関数と定義し，カメラの位置姿勢，空間点位置をパラメータとして，非線形最小二乗法により最適化を行う手法である．バンドル調整に関する理論から実装までの詳しい解説は，岡谷のチュートリアル[79] や，Triggs らの解説[75]，Snavely の博士論文[68] を参照されたい．

バンドル調整において，再投影誤差の最小化問題は，非線形最小二乗問題として定式化され，反復解法であるガウス・ニュートン法やレベンバーグ・マーカート法が用いられる．ある空間点の再投影誤差はそれ以外の点の空間座標には依存しない，さらに，その画像に関連しないカメラの位置姿勢にも依存しない．このことから，計算の過程で用いるヤコビ行列 $\mathtt{J}$ やヘッセ行列の近似 $\mathtt{J}^\top\mathtt{J}$ が非常に疎な行列になるため，パラメータの並びを考慮し，シューア補行列トリックを用いることで，計算効率を大幅に向上できる[75][79]．現在広く利用されているバンドル調整のライブラリとして， Lourakis による Sparse Bundle Adjustment(SBA)[38] がある．SBA ライブラリの詳細については[37] を参照されたい．本ライブラリの投影関数を変更することで，全方位カメラ画像，魚眼レンズカメラ画像での SBA を行うことも容易である[25]．

SfM における refinement の手法として，SBA は最も現実的な方法であり，数百，千枚の画像に対しては，十分に効率的である[1]．しかしながら，数万というオーダーの画像に対してバンドル調整を行うためには，さらなる工夫が必要であり，現在，復元問題において再注

⋆1 縮退の一例として，空間の点が全て平面上にある場合，８点アルゴリズムを用いて F 行列は推定できない．

© 2011 Information Processing Society of Japan

目されているテーマのひとつである．

Ni らは[50]，大きなサイズの行列演算の反復を避けるために，分割統治によるバンドル調整を提案した．彼らが提案した out-of-core bundle adjustment の特徴的な点は以下の部分にある．(1) 観測データを submap に分割し，submap の中で，他の submap との干渉のない内部パラメータと干渉を持つ相互パラメータに分解する．(2) Submap をパラレルに最適化，各 submap からの相互パラメータ部分を抽出，キャッシュし線形近似を行い更新する．(3) 相互パラメータの更新情報を内部パラメータに伝搬する．(2)，(3) を繰り返すことで，大きなデータに対する演算を避け，メモリの消費量，計算量をともに抑えることができるものの，はじめに充分に良い submap 分割を行う必要がある．

大量の画像を扱う場合，シューア補行列を用いたとしても，その行列計算のコストが大きい．ガウス・ニュートン反復はどのみち近似最適化であるため，各ステップにおける厳密解を求めず，共役勾配法 (conjugate gradient) を用いたアプローチが提案されている[2]．共役勾配法を用いる場合，正規方程式 $\mathtt{A}\mathbf{x} = \mathbf{b}$ における $\mathtt{A}$ 条件数が大きいと収束が悪いことが知られている．バンドル調整において，$\mathtt{A} = \mathtt{J}^\top\mathtt{J}$ であり，常に条件数が高い．そこで，上記方程式の両辺に，Block Jacobi preconditioner 等の変換行列 (preconditioner) を作用させることで改善を図る[1]．さらに，同会議で発表された Byrod らの手法[8] では，共役勾配法において $\mathtt{J}^\top\mathtt{J}$ の計算を回避する定式化を行なっている．また，共役勾配法ベースのバンドル調整の効果は，シーン構成に大きく依存することが報告されており[1)8)]，シーケンシャルな画像列 (street view など) や，中小規模の問題では，SBA の利用が推奨されている．

1.4 大量画像復元のアプローチ

カメラの位置姿勢と空間点の復元は，正射影カメラの場合，因子分解法 (Factorization)[72] による閉形式解が存在する．因子分解法の paraperspective camera へ拡張[57]，射影カメラへの拡張[11)70)] なども存在するあるが， perspective カメラでは，代数的距離の最小化であり直接的な幾何学的解釈が存在しない．また，反復的解法が存在するが[72)23)15)]，観測点に欠損データが存在する場合 (トラッキングが途切れている場合) は，閉形式解が存在しない．ロバストなエラー関数を因子分解法のスキームに取り入れるのは困難であるため，カメラ運動や撮影条件の事前情報がなく，ノイズ，誤対応を多く含みやすい大規模データからの SfM には，不向きなようである．

そこで，Bundler では，画像ペアのマッチン情報を元に，画像の接続関係を示す image connectivity graph を作成し，その情報を元に，初期復元ペアを選択し，５点アルゴリズムによりカメラ位置姿勢・空間点を復元，画像を追加，復元，バンドル調整を繰り返す．

上記の復元アプローチでは，初期復元ペアの選択が非常に重要である．Bundler では，初期復元ペアとして，epipolar geometry マッチングによって得られる対応点の数が１００点以上かつ，homography マッチング[23] によって得られる対応点の数が最も少ない，画像ペアを選択している．このことにより，幾何学的縮退の起きている悪条件な画像ペアを選択するのを防いでいる．

Image connectivity graph は，単純に画像ペア間に track が存在するか否かをエッジとするグラフであるが，他の情報をエッジの重みとするグラフも構成可能である．Gherardi らは，エッジの重みとして geometric robust information criteria(GRIC)[73] を用いた image graph を構成している[19]．さらに，良い初期復元の重要性から image graph からバランスの良いデンドログラムを構成し，それに基づいた階層的な復元戦略と[19]， autocalibration による内部パラメータ推定とその推定モデルのテスト[20] を行なっている．

ここで，flickr などから得た未整列画像には，duplicated image や nearly duplicated image など，復元において冗長な画像が多く含まれる．シーン全体を効率良く復元するために，まず，image graph の部分グラフとして構成される skeletal set を作成し，データセットのサイズを削減してから復元を行う．skeletal set に含まれなかった画像は後で追加する．ここで，シーンを表現する上で冗長な画像を省きながら，精度を保ちつつ，グラフ全体としての接続関係を保つような skeletal set の構築が必要である．Snavely らは，２枚のカメラ位置から算出した共分散行列のトレース[66] をエッジにもつ，image graph を作成し，minimum connected dominating[⋆1] と t-spanner を用い，skeletal set を構成した．このアイデアは彼らの論文の発展型,[2] にも取り入れられている．

Geometric verification を行なったとしても，画像ペアマッチングで得られる対応点に誤対応 (ミスマッチ) が存在しうる．例えば，エピポーラ線上に乗っている対応点や，エピポール付近の点は対応点を許容する閾値の大きさに関係なく誤認されやすい．多くの誤対応は，２枚の画像間の対応だけでなく，複数にわたる対応 (track) を生成する段階で除去されたり，DLT アルゴリズムでカメラ位置姿勢を推定する段階で除去される．

しかしながら，ミスマッチを含む画像ペア，ミスマッチから計算された間違ったカメラ位置姿勢を加えることで，その後の復元が破綻する可能性は常に残る．したがって，より信頼性の高い image connectivity graph や image graph の生成が重要になる．グラフに内在する loop を利用することで，そのようなミスマッチを除去することが可能である．ループ

⋆1 論文中では maximum leaf spanning tree と呼ばれている．

© 2011 Information Processing Society of Japan

グラフのあるノードからスタートして，画像ペアマッチングによって得られたカメラ間の回転や homography 行列をエッジに沿ってかけ合わせると，スタートノードに戻ったときに，誤差がなければその行列は閉じている (単位行列またはその定数倍) はずである[41]．現実的に，存在する全てのループ拘束を実際に検証するのは困難であるため，Zach らはローカルグラフのみで検証するアルゴリズムを提案，ビルの窓枠等の repetitive pattern を多く含む，非常に曖昧性の高い画像を含むデータセットにおいて，その効果を証明している[78]．

1.5 画像認識手法との融合

マッチングを行うペア数は quadratic であるため，ペア数の削減を行うために，画像を何らかの形で符号化し，実際のマッチングを行う前に，近似的な画像の類似度を算出する．そして，類似度の高い画像のみマッチングを行う．このようなアイデアは，Schaffalitzky-Zisserman の ECCV2002 論文[62] にすでにに含まれており，数十枚の画像というサイズでの実験に成功している．

類似画像検索手法の洗練に伴い，GIST[56] や visual words and vocabulary[63] を用いた画像の分類手法が，３次元復元においても広く用いられるようになった[35][25][2]．

Li らは，上記の画像類似度による画像データセットのクラスタリングを行うことで，4 万枚の入力画像からランドマークの復元を効率よく行う方法を提案した[35]．システムの特徴をまとめると以下のとおりである．

(1) 全ての画像を GIST descriptor[56] で表現，k-means clustering によって分割
(2) 各クラスターから代表的な N(=8) 枚の画像を選択，その中でマッチングを行い inlier match の数が最多のものを iconic image として選択
(3) 各クラスターの iconic image から，Nister-Stewenius の vocabulary tree[54] を用い iconic scene graph を作成 (image graph に相当)
(4) 画像に付加されている tag 情報を用いたフィルタリングの後，SfM を行う．

同著者らによる発展版が[17] であり，現段階における最新アルゴリズムの集大成になっており，１台の PC のみを用いて，２４時間で 6.4 万枚の画像の SfM と dense reconstruction を行なっている．

同様のアプローチとしては，Havlena らによる Randomized SfM がある[25]．

(1) Visual words and vocabulary[63] を用い，各画像を tf-idf vector として表現.
(2) tf-idf vector の内積により算出されるスコアを類似度として image similarity graph を作成.
(3) スコアの高いものから３組の画像を選択し復元 (atomic model)，十分数の atomic model を生成した後，相似変換を介してマージ.

tf-idf vector の内積により算出される高速な類似度計算，３視点からの復元により安定な atomic model を生成している．復元全体が初期ペアの依存することを回避するという特徴を持つ．さらに，この研究の発展研究として，image similarity graph から minimum connected dominating set(CDS) を作成し，復元を行うことを提案した．Skeletal graph は実際にマッチングを行い作成した image graph から生成されていることを考慮すると，image similarity graph からの生成する CDS は skeletal graph の近似であるが，充分に良い近似であることを実証している．さらに，SfM の各処理をタスクとして捉え，その優先順位に応じて順に (簡単な問題から先に) 処理を行うことで，復元処理が停滞するのを防いでいる．

Agarwal らは，Flickr に ”Rome ”または”Roma” とタグ付けられた 15 万枚の画像から２４時間で２６箇所の主要なランドマークの復元に成功している[2]．システムの各要素は標準的なものであるが，６２ノードで構成されるクラスター PC (コア数は４９６) での分散処理による劇的な処理速度の向上を測った．クラスターシステムを用いた復元では，復元タスクの分散とロードバランスのメンテナンスという異なった次元での挑戦になっている．

2. 実時間 SfM と Visual SLAM

Visual SLAM(Simultaneous Localization And Mapping) とは，ロボティクスの分野にルーツを持つ技術で，SfM 同様，多視点画像を用いて対象の 3 次元形状とカメラの運動を推定するが，特に実時間性および因果性の制約下でこれを行うものである．ロボティクスでは古くから SLAM，すなわち未知の 3 次元空間を移動ロボットが探索する際，自己の位置を推定しつつ，同時に空間の地図を構築してゆく方法が研究されていた．そこではレーザレンジセンサなどが使われることが多かったが，その SLAM の技術を，センサにカメラ（のみ）に変えた場合に応用したのが Visual SLAM と言える．

SfM は，Visual SLAM よりも長い歴史を持つが，当初想定されていた応用の性質から，カメラ運動の推定よりも 3 次元形状の推定に関心があることが多く，また実時間性や因果性が求められることはなかった．しかしその後，CG と連携した映像制作や仮想現実など，カメラ運動の推定そのものが目的となる応用が開拓され，特に実時間性を前提とする SfM と Visual SLAM は，現在では同一視されるようになった．

2.1 問題の概要

目的は，カメラが空間を制約なく移動し，画像を時々刻々撮影するとき，その画像列から

© 2011 Information Processing Society of Japan

カメラの運動と周囲環境の 3 次元形状を，実時間で因果的に推定することである．カメラにはステレオカメラなど剛体接合された複数カメラを使う場合もあるが，単眼でもよい．以下では主に単眼カメラを扱う場合を考える．複数カメラを使うと，それが静止状態であっても 3 次元復元が可能となり，レーザレンジセンサなど，奥行きが計測可能なセンサを使う場合と似た問題になる．

カメラの運動と環境の 3 次元形状を推定する基本的な原理は，前節で解説のあった未整列画像を対象とした SfM と同じである．一つの違いは，画像間での特徴量の対応付けがずっと容易になることである．特徴量には主に点，場合によっては直線[82)–84)] を用いるが，カメラ運動の連続性から画像も連続性があるので，画像間での変動が小さい．より一般的には，カメラの運動情報から特徴量の画像上の位置をある程度予測できるため，画像上で探索すべき領域を小さく制約できる．これは対応探索の精度を向上し，計算時間も低減する．

異なる視点間の特徴量の対応が十分な数あれば，カメラの姿勢および特徴量の 3 次元位置は推定できる．ただし実時間性の制約から，これは限られた計算量で実行する必要がある．また過去の画像しか推定に利用できない因果性は，因果性に制約されないオフラインの SfM に比べると，精度の維持をいくぶん困難にする．さらに，復元対象とする空間の規模を大きくしたいという矛盾する要求が加わる．

2.2 2 つのアプローチ：フィルタリングと部分的バンドル調整

これらの課題を解決するために，2 つの相反するアプローチが検討されてきた．毎フレームごとのカメラの運動推定（トラッキング）と，環境の 3 次元形状復元（地図生成，マッピング）を同時に結びつけて行う方法と，逆にこれらを切り分け，並行に行う方法の 2 つである．計算方法としては，前者は確率的フィルタリングを用いた逐次計算，後者はバンドル調整にそれぞれ対応する．この視座[85)] から，以下では，フィルタリングに基づく方法とバンドル調整に基づく方法の 2 つに分けて，既存研究を説明する．

2.3 フィルタリングアプローチ

観測 z_k が時系列で $k = 1, \ldots$ と与えられるとき，過去の観測 $\{z_1, \ldots, z_k\}$ を余すことなく使って，現在時刻の状態 x_k をなるべく高精度に推定したい．これを可能にするのが，カルマンフィルタに代表される再帰的なベイズ推定に基づく，確率的フィルタである．これは，過去の観測が与えられたときの現在の状態の事後分布 $p(x_k \,|\, z_k, \ldots, z_1)$ を，再帰的に推定する．

そのためには 2 つの要素，観測の確率モデル $p(z_k \,|\, x_k)$ と，直前の状態から現在の状態を確率的に予測する $p(x_k \,|\, x_{k-1})$ を指定する必要がある．今の問題では，状態 x_k は現在時刻のカメラの姿勢と，空間の点すべての 3 次元座標の 2 つを並べたベクトルになる．観測モデルは，x_k すなわち現在のカメラ姿勢と空間の各点の位置を与えたとき，画像上の点の投影位置を表現する（投影の式と特徴点の位置誤差の確率モデルを合わせたもの）．また，予測のモデルには，カメラの運動モデル（例えば等速直線運動など）を使うことができる．

これら観測と予測のモデルが，変数について線形かつガウス性を有する場合，上述の各分布もまたガウス分布となり，その平均と分散のみを再帰的に計算すればよい（カルマンフィルタ）．しかし今の問題では，主に観測モデルの非線形性からそうはならない．一つの方法は，変数の位置でこれを線形近似し，再帰計算を繰り返す近似手法，すなわち拡張カルマンフィルタ (EKF/Extended Kalman Filter) を使うことである．

2.3.1 初期の研究

動画像を使った色々な幾何学情報の推定にこのようなフィルタを用いる方法の研究は 80 年代にさかのぼる． Gennery は既知物体の位置・姿勢をカルマンフィルタに似たフィルタを使って実時間推定する方法を述べている[86),87)]．同時期の同様の研究に Dickmanns[88)] らがある．これらは 3 次元形状を既知としており，それとカメラ運動を同時に推定するものではなかった．

対象の 3 次元形状とカメラ運動の両方を同時に推定する方法の最初の一つが，Harris-Pike[89)] である．そこでは，カメラ運動は非線形最適化で求め，特徴点の位置推定にカルマンフィルタ（線形のもの）を用いている．特徴点の位置をユークリッド座標にて表現すると，その分布はガウス性からかい離するとして，画像座標＋視差を用いてこれを表現していることや，カメラ運動の最適化に特徴点の位置の共分散行列を利用して高精度化を図ったことなど，近年の方法にも通じる点が多い．

カメラ運動と点の位置をフィルタを用いて同時推定する，現代的な方法の最初は，Broida らの研究[90)] である．単眼カメラで追跡した特徴点を使って，EKF によりパラメータの因果的推定を行った．これは Azarbayejani-Pentland[91)] で拡張され，焦点距離も同時推定された．また Chiuso-Satto らの MFm[92)] では，特徴点の消失に伴う状態変数の削除と，新たな特徴点の追加に伴う状態変数の追加の方法が議論された．

McLauchlan-Murray[93)] は，VSDF(Variable State Dimension Filter) と呼ぶ，状態変数の追加と削除が柔軟に行える再帰計算の枠組みを示している．これは，バンドル調整の解説[94)] にも記載されており，状態変数をすべて最適化する完全バンドル調整と，EKF の間を橋渡しするものとなっている．その後，McLauchlan[95)] にて，SfM の再帰的計算（フィルタリング）とバッチ計算（バンドル調整）の有機的関係について述べている．

© 2011 Information Processing Society of Japan

2.3.2 Davison の MonoSLAM

Visual SLAM と呼んでいるものを最初に具現化して見せたのが Davison の MonoSLAM[96),97)] である．その名の通り単眼カメラを利用し，カメラの自己位置推定と周囲環境の 3 次元地図構築を同時に実時間で行うという命題をはっきりと示し，それを実現した点で画期的であった．この研究は，ロボティクスで確立された SLAM 技術を，単眼の SfM 問題に持ち込んだものと見なせる（同じ著者の以前の研究に，能動ステレオカメラを用いて同様に SLAM を行う方法[98)] がある）．特徴点（の 3 次元位置）をランドマークと呼び，後述のように，視野を外れた後の再認識を考えたことは，まさにロボティクス的発想だったと言える．

MonoSLAM は，校正済みカメラの現在の姿勢とランドマークの空間座標を状態変数 x にとり，標準的な EKF を使って，x の平均および共分散行列をビデオレートで更新する．この点では，地図構築か形状復元かという文脈の多少の違いを除けば，上述の従来研究[90),91)] などとそう違わない．従来研究との明確な違いは，一度 3 次元位置を推定した特徴点の「再認識」を可能にした点である．つまり，特徴点がカメラの運動に伴って一度視野の外へ出た後，再び視野内に戻ってくることがあれば，その点を同じ点として認識できる．これによってカメラの姿勢推定および地図構築の両方を高精度化した．なお従来の SfM では，画像系列内で追跡した特徴点の「連続軌跡」を観測とし，一度視野から外れるなどによりいったん消滅した点はそれきりだった．そのような点が再度視野内に現れても，同じ点と見なすような発想や仕組みは一般的でなかった．

MonoSLAM では既存のランドマークを認識するのに，その画像における投影位置を，現在のカメラ姿勢の推定値を元に予測し，サーチ領域を限定した上で画像パッチの相関（正規化 SSD を使用）によって求めている．各ランドマークはその位置の不確かさをガウス分布として保持しているため，カメラ姿勢の確率的推定と合わせて，この予測は画像上の 2 次元ガウス分布となる（したがって画像上のサーチ領域は楕円になる）．

カメラから見えるランドマークの数が閾値を下回ると，新たなランドマークを追加する．特徴点の検出には Shi-Tomasi の方法[99)] を使っている．また，認識に一定の割合で失敗するランドマーク（見えているはずなのに対応付けに失敗する）は削除も行う．

初期の MonoSLAM では，ランドマークの位置を状態変数として表現するのに，その 3 次元座標そのものを使っていた．しかしながらこの表現の下では，観測モデル（空間の点から画像座標への投影変換）の非線形性が強いため，特にランドマークの追加時，その 3 次元位置を初期化するのに特段の工夫が必要であった．新たなランドマーク追加時，そのランドマークを見る視点の変化が小さい（基線長が短い）ため，点の奥行きの分布は，カメラに近いところから無限遠まで，ある点をはさんで非対称なものとなるはずである．これを単純なガウス分布で近似すると，点の位置の不確かさを誤って小さく見積もることになる．そのため，新しいランドマークを一定の間，古いランドマークとは別に扱うようにし，視点の変動が十分起こったところで，EKF に取り込むようにしていた（遅延初期化）．その後，Civera-Davison-Montiel[100)] で，ランドマークの位置を（特定の視点で見た）奥行きの逆数を用いて表現することで，この問題を回避する方法が示された（非遅延初期化）．

2.3.3 EKF-SLAM の問題点

EKF を使った SLAM にはいくつかの問題点がある．一つは，特徴点（ランドマーク）の増加に従って計算量が爆発的に増えてしまうことである．これは，カメラの姿勢は最新のもののみ状態変数に取りこみ，過去のすべての姿勢を周辺化するフィルタリングの本質にその原因がある．MonoSLAM では，ランドマークの数と同じオーダの共分散行列を毎時刻更新する．この行列は初期状態こそ疎で計算量は小さいが，上の周辺化にともなって急速に密行列へと変化（fill-in）し，更新のための計算量が増加する．このことから，MonoSLAM はせいぜい 100 程度のオーダのランドマークを扱うのがやっとであり，広いスペースを長時間にわたって探索することには向かない．

EKF のもうひとつの問題点は，観測モデルの非線形性に伴うものである．EKF は観測モデルを都度，線形近似し，また状態変数の事後分布をガウス分布で近似する方法であるから，推定精度を維持するには，これらの近似精度が担保されることが必要条件である．この条件が満たされない場合，特に観測モデルの非線形性が無視できないとき，EKF はしばしばいわゆる一致性 (consistency) を失い，具体的には状態変数の不確かさ（共分散）を実際よりいつも小さく見積もってしまうことが知られている[101)]．文献[102)] では，EKF-SLAM で実際に一致性が失われることと，その原因の考察が示されている．

以上の EKF-SLAM の問題はロボティクスの分野でよく認識されており，これを解決するために FastSLAM という方法が提案されている[103)]．この FastSLAM はラオブラックウェル化パーティクルフィルタ（Rao-Blackwellized PF）に基づいている．Eade-Drummond は，この FastSLAM に基づいた単眼 SLAM のシステムを示し，EKF ベースの MonoSLAM に対する優位性を述べた[104)]．ただし，EKF-SLAM より性能は改善するものの，FastSLAM もパーティクル数が不十分な場合にやはり一致性に欠けることが指摘されている[105)]．

2.3.4 Eade-Drummond の方法

実時間 SfM/Visual SLAM にフィルタを用いたときの統計的な不一致性，すなわち推定精度の低下が，非線形な観測モデルの線形化の際に生じる誤差を原因とするのであれば，こ

© 2011 Information Processing Society of Japan

れをなるべく小さくすることがその解決につながる．このことと，広い空間を自由に探索できることとを両立する一つの方法が，Eade-Drummond[106] に示された．その方法とは，空間を複数の部分空間に分割して表現するもので，その部分空間ごとに局所座標系を与えて，その内部でのみフィルタリングを実施する．

フィルタリングは一つの局所座標系内のみで実行されるので，観測モデルの線形化の影響はその局所座標系内に限定され，全体には波及しない．その際，線形近似がなるべく精度を維持できるように，ランドマークのパラメータ表現を文献[100] 同様，ユークリッド座標ではなく，奥行きの逆数によって与えている．

局所座標系間は相似変換で結ばれる．局所座標系をノードとすると，相似変換はノード間のエッジを与え，全体の構造は一つのグラフで表現される．この相似変換は，それが結ぶノード（局所座標系）間で共通するランドマークを使って推定される．より詳細には，2 種類の制約を用いて最適化を行う．一つは，共通ランドマークの観測が与える相似変換の制約で，もう一つは，グラフ内の閉路を一周したときの恒等性である．具体的には，グラフに対しスパニング木を見つけ，これに含まれるエッジについては前者の制約を，含まれないエッジは（グラフ内で閉路を形成するものとなるので）後者のそれを，それぞれ尤度で表現しその積を最大化して行う．なお，この最適化はグラフ全体で大域的に行うが，対象となるのはエッジ（=局所座標系間の相似変換）のみであって，各局所座標系内のランドマークは全く変更されない．

また，既探索の空間をカメラが運動するときは，現在どの局所座標系にカメラがあるかを判断する必要がある．また，未探索空間に入るときは，新たに局所座標系を作り出す必要があり，それをいつ行うかを決める必要がある．これらの判断に，観測のランドマークパラメータに関する 2 階微分（ヘッセ行列）を用いるヒューリスティックな方法を使っている．これは，観測モデルが線形に近ければ，一般に 2 階微分は小さくなるはずだという考えに基づいてる．

以上の処理の手順は以下のようになる．

1. 既知のランドマークの最新の画像上での対応を探す．現在アクティブなノードとその周辺ノードにおける状態の推定値から，現画像上の対応点位置を予測し，探索範囲を狭める．
2. アクティブなノードを選択する．上述のヘッセ行列を用いた方法により，ステップ 1 で得た観測に対する観測モデルが最も線形に近くなるノードを選ぶ．もし適当なノードがなければ，新たに生成する．
3. アクティブなノードにおける状態変数を，EKF に似たフィルタリングにより更新する．新たなランドマークを FAST[107] によって検出する．
4. グラフを更新する．共通のランドマークを持つノードどうしをエッジで結び，相似変換を計算する．グラフ内の閉路に対し，各エッジ（相似変換）をガウスニュートン法で最適化する．

同論文では実データを用いて，本方法が，MonoSLAM のような EKF-SLAM および Fast-SLAM ベースの方法[104] よりも高精度であることを，オフラインの完全バンドル調整の推定結果と比較して示している．

2.4 バンドル調整によるアプローチ

フィルタリングとは別のアプローチがバンドル調整に基づく方法である．フィルタリングは，逐次的に得られる観測を基本的にすべて用いて推定の精度を向上しようとするが，反面，最新のカメラ姿勢のみを状態変数として保持する場合，過去のカメラ姿勢を周辺化して系から（表面的に）削除するため，ランドマーク数に対して計算量が爆発する問題や，本来非線形なシステムを線形化することによる近似誤差が逐次計算の過程で蓄積し，最終的に大きな誤差になってしまう問題があった．

3 次元復元の精度を考えれば，画像をすべて使う完全なバンドル調整が最良であることは自明である．問題は，逐次的に得られる画像すべてを用いてバンドル調整を行ったのでは，計算量が大きくなりすぎることである．実時間性を考慮したアプリケーションでは，それではほとんど現実性がない．そこで，時々刻々得られる画像列から，新しい方から何枚かの画像を抜き出し，その画像についてのみカメラ姿勢を推定するようにバンドル調整を行い，計算量を抑える方法が考えられるに至った．

2.4.1 初期の方法：オフラインの SfM

まず，運動カメラで撮影した連続画像を用いた SfM の方法について述べておく．これは，ここでの問題設定とは異なり，オフラインの復元を目指したものだが，後の研究に少なからず影響を与えている．

同一シーンの連続画像を使う SfM では，隣接画像間での局所的な 3 次元復元を延長するシーケンシャルな方法が当初一般的であった．これに対し，Fitzgibbon-Zisserman[108] で，それに代わる階層的なアプローチが提唱された．ビデオ画像の連続するすべての 3 枚組みの画像に対し，特徴点を追跡し 3 重線形テンソル（trifocal tensor）を推定する．さらに 3 枚組の重複の関係を利用し，射影復元した特徴点の空間座標とカメラの姿勢を同じ空間での表現に直し，これを繰り返すことで全画像系列に対する，一つの 3 次元復元結果を得る．

© 2011 Information Processing Society of Japan

Nistér[109] はこれを拡張し，時間的に連続する 3 枚の画像が復元に適しているわけではないことから，Torr ら[110] の方法を用いて適切な 3 枚の画像を系列から抜き出す方法を示した．画像特徴として，点と直線が利用された．

上の方法は，復元結果をより高精度化するのにバンドル調整を最後に実行していた．そこで次に，このバンドル調整の過程そのものを効率化しようとする研究がなされた．初期の研究に，Shum らの階層的バンドル調整[111] や Zhang の最新画像 3 枚を使った部分的バンドル調整[112] がある．前者は，オフラインのバンドル調整を効率化する方法だが，後者は逐次 SfM の実時間実行を対象としたものである．また，Engels-Stewénius-Nistér[113] では，実時間性を考慮した SfM でのバンドル調整の性能を評価している．過去何視点のカメラ姿勢分まで最適化の対象とするかと，最適化の反復計算を何回行うかに応じて，結果がいかに変化するかが調べられている．

画像系列からある枚数の画像を選ぶとき，Zhang[112]（や Fitzgibbon-Zisserman[108] のように）最新のものから連続した画像を選ぶのではなく，Nistér[109] がそうしたように，何枚かおきに画像（=キーフレーム）を選んだ方が，全体の精度が高くなりそうである．キーフレームという考え方は Shum らの研究[111] にもあったが，これを実時間 SfM で最初に実現したのが，後述する Mouragnon らの研究[114),115] である．この研究は PTAM[116] にも影響を与えている．

2.4.2 Nistér の Visual Odometry

Nistér らは，実時間・因果性の制約の中で，画像系列の一部に対する最適化を逐次的に繰り返し，高精度に SfM を実行できることを示した[117]．他にセンサを用いず単眼カメラのみでカメラ運動を高精度に推定できることから，Visual Odometry(VO) と呼んだ．後述のように，その方法はバンドル調整とは言えないが，精度と計算量をうまくバランスして見せたことは後の研究に大きな影響を与えている．

校正済みの単眼カメラを対象としたアルゴリズムは次の通りである．

1. 一定の長さの画像列上で特徴点を追跡する．具体的には，Harris コーナーを毎フレーム多数検出し，フレーム間でマッチングを行っている（KLT などとは対照的）．この系列内の 3 枚の画像でのカメラの位置・姿勢を，5 点アルゴリズム[118] およびプリエンプティブ RANSAC[51] を用いてロバストに求め，非線形最小化（詳述されていないが恐らくカメラ運動についてのみの最小化）を行う．
2. 追跡した各特徴点軌跡に対し，最初と最後の観測を元にその 3 次元位置を文献[119] の閉形式解法で計算する．復元結果のスケールを再度プリエンプティブ RANSAC を用いて決定し，過去の復元結果と同一の座標系上に定める．
3. さらに何枚かの画像列上に追跡を延長する．3 次元位置が既知の特徴点を元に 3 点アルゴリズム[120] およびプリエンプティブ RANSAC を用いてカメラの位置・姿勢を求め，非線形最小化を行う．
4. 特徴点の 3 次元位置を再計算する．ステップ 3 と本ステップを何回か繰り返す．
5. ステップ 1 から何度か繰り返す．
6. 推定結果をリセットし，ステップ 1 から再度新たに開始する．

なお，この論文[117] には，ステレオカメラを用いた場合のアルゴリズムも示されている．

以上から分かるように，カメラの運動と特徴点の位置を分離して，交互に定めている．この点でこの方法はバンドル調整とは言えず，resection-intersection[94]（点を固定しカメラ姿勢を推定する resection とカメラ姿勢を固定し点を推定する intersection の交互反復）の一種と見なせる．方法は高い性能を示したが，その後の研究[113] では，そのような場当たり的な解法は良くなく，バンドル調整を行うべきだとある．

2.4.3 Mouragnon らの方法

Mouragnon らは，単眼カメラから得られる画像系列の一部にバンドル調整を適用することで，精度を維持しつつ，逐次的な 3 次元復元を行う方法を示した[114),115]．そこでは，最新の画像からさかのぼって何枚かの画像のみを使って，部分的なバンドル調整を行うことと，その際，時間的に連続した画像を使用するのではなく，なるべく時空間的に離れた画像をキーフレームに選定し，精度と計算時間の両立を目指すことがポイントである．

類似のキーフレームの概念は Shum らの階層的バンドル調整[111] にも見られるが，これはオフラインの SfM の効率化のためのものである．逐次復元をターゲットとした類似研究に Zhang[112] があるが，これは最新の画像 3 枚を使ってバンドル調整を行うもので（さらに点の位置を推定せずカメラ運動のみを推定することでさらなる効率化を図る方法），キーフレームの概念はなかった．

なお Mouragnon らの研究では滑らかに変化する画像を想定しており，特徴点に Harris コーナーを使用，その追跡を相関ベースで行っている．(つまり，MonoSLAM[96] のような特徴点の再認識機能はない.）バンドル調整は，画像系列からある基準で選択したキーフレームの画像群にたいしてのみ行う．最適化されるパラメータは n 個のキーフレームでのカメラの姿勢とそれら画像上の点の 3 次元位置であるが，観測は（n より多い）N 枚の画像を使う．なお，キーフレームの間ではカメラ運動のみを，通常のやり方（P3P アルゴリズム[120] と RANSAC を用いたロバスト推定の後，再投影誤差最小化）で求める．

© 2011 Information Processing Society of Japan

キーフレームの選択基準，つまり，いつキーフレームを生成するかが問題となる．彼らは，最新の画像と最後のキーフレーム間で共通する特徴点の数の減少と，最新の画像でのカメラの姿勢の精度の低下の 2 つをそのトリガーに選んでいる．前者は共通する特徴点数が 400 個を下回った瞬間，後者は再投影誤差最小化時のカメラ姿勢パラメータ（6 自由度）のヘッセ行列の逆数（共分散行列の推定値となる）を用いて算出する．

アルゴリズムは次のようにまとめられる．

1. 初期化を行う．共通する特徴点の個数が一定以下になる直前の画像をキーフレームとして 3 枚選択し，それらにおけるカメラの相対姿勢変化および特徴点の 3 次元位置を求め，基準座標系を定める．
2. 最新の画像と最後のキーフレーム間のマッチングを行い，カメラの姿勢およびその精度を求める．もし新たなキーフレームを生成する必要があるかどうかを検証する．もし必要がなければこのステップを繰り返す．
3. 直前の画像を新しいキーフレームとして追加する．新たな特徴点の 3 次元位置を求め，部分バンドル調整を行う．2 へ．

この方法では，毎フレームのカメラ運動の推定（ステップ 2）は，既知の特徴点の座標を固定し，カメラの姿勢のみを観測に基づいて最適化して行う．この考え方は，次に述べる PTAM と基本的に同一である．

2.4.4 Klein-Murray の PTAM

Klein-Murray の PTAM(Parallel Tracking and Mapping)[116] は，もっとも成功した単眼 SLAM のシステムと言える．もっぱら小さな閉空間での AR のために設計されており，広大な空間を探索する SLAM には明らかに不向きな部分はあるが，上述したような過去の研究の長所をうまく取り入れて高度なシステムを構築した点で，画期的であった．

PTAM の最大の特徴は，その名の示す通り，トラッキング（=毎フレームのカメラ姿勢推定）と 3 次元復元（地図生成）を分離したことである．これは，上述の MonoSLAM[96] や，Eade-Drummond の方法[104),106)] のようなフィルタリングの方法と対照的であり，画像からキーフレームを選んでバンドル調整を行うという概念が確立された．トラッキングと 3 次元復元を分離する考えの背景には，PC の CPU がマルチコア化され，複数スレッドを物理的に並列に実行できるようになったこともある．

トラッキングは，3 次元復元されたマップ点（=MonoSLAM[96] でいうランドマーク/特徴点と同じ）の情報を元に行う．このマップ点は，キーフレーム上で検出され，他のキーフレームと対応付けられて 3 次元復元されたもので，この処理は後述の地図生成スレッドにおいて実行される．各マップ点には，その 3 次元座標と物体表面の法線ベクトルの各情報が与えられ，さらに特定のキーフレームの画像の部分領域への参照を与えることでその見えが与えられる．これを用いてトラッキングは，MonoSLAM[96] に類似した次の手順で行われる．

1. カメラから画像を得る．その時点のカメラ姿勢を以前の推定結果を使って予測する．
2. そのカメラ姿勢に従って，マップ点を画像上に投影し，対応を探す．
3. 見つかった対応関係からカメラ姿勢を推定する．

なお，特徴点の検出自体は FAST-10[107] を用い，画像の解像度を 4 段階に変えた画像ピラミッドの各階層に対して実行する．対応探索には SSD が用いられる．上のトラッキングにおけるステップ 2 および 3 は，2 段階で行われる．まず，この画像ピラミッドの最上層（もっとも粗い解像度）にある特徴点を 50 個用いて行い，次に全階層の特徴点 1000 個を使って行う．

トラッキングと並行して行われる 3 次元復元（地図生成）は，次のような手順で行われる．まず，キーフレームの生成は，次のような条件が満たされた場合に行う．

a) トラッキングが一定以上の精度で行えていること．対応付けができたマップ点の数によって評価される．
b) 最後にキーフ レームが追加されてから 20 フレーム以上経過していること．
c) 最も近いキーフレームから一定以上離れていること．距離はキーフレームのカメラ位置に対して計算する．この閾値は，復元したマップ点群へのカメラからの距離に応じて定められる．

なおキーフレームが生成されたとき，新しいマップ点を追加するとともに，最近傍にあるキーフレームとの間でエピポーラ条件を用いてそれらの対応を探索し，3 次元位置を定める．なお定まった 3 次元位置は後に随時最適化されて，計算し直される．

キーフレームが生成されないとき，3 次元復元のスレッドは，部分的なバンドル調整あるいは，完全なバンドル調整を実行する．部分的なバンドル調整とは，キーフレームの部分集合（例えば 3 つ程度）に対し，関連するパラメータ（カメラ姿勢およびマップ点の位置）を最適化する．基本的な考え方は，Mouragnon らの方法[114] を踏襲している．完全なバンドル調整とは，全キーフレームのカメラ姿勢および全マップ点の位置を最適化する．いずれのバンドル調整でも，M 推定の考え方でロバスト化した再投影誤差を最小化している．

このようにバンドル調整が 2 段階に運用されるのは，完全バンドル調整のみとすると，キーフレーム数が大きくなってきたときに計算量が大きくなりすぎ，収束まで時間がかかりすぎることによる．新しいキーフレーム周辺の 3 次元復元の精度を確保するのがまずは大

© 2011 Information Processing Society of Japan

事なので，そのための仕掛けである．なお，いずれのバンドル調整もキーフレームが生成されると中断され，上述のキーフレーム生成時の処理を優先するようになっている．つまり，時間に余裕があるときのみ，バンドル調整が実行され，復元された地図が高精度化してゆくということになる．

その後 PTAM は，エッジ情報を特徴量として用いることで，より俊敏なカメラ運動に対応できるように拡張され[121]，またスマートフォンのカメラのように性能の低いイメージングシステムで PTAM を行う方法[122] も提案された．

2.5 まとめと議論

以上，フィルタを使うアプローチとバンドル調整に基づくアプローチに 2 分し，単眼カメラを使った実時間 SfM/SLAM の各研究を説明した．2 つのアプローチのどちらが良いかだが，Strasdat-Davison[85] に比較がある．そこでは，両アプローチの計算量と推定精度をシミュレーション実験によって評価している．そこでの結論は，ごくわずかな時間のうち計算を終えなければならない場合を除き，バンドル調整の側に分があるというものであった．ただし，バンドル調整（=バッチ最小二乗）とフィルタリング（=再帰的最小二乗）の関係[93),95)] を考えると，この研究だけで結論するのは早尚であるように感じられる．

なお，フィルタリングではなく，スムージングを SLAM に適用する考え方（iSAM(Smoothing and Mapping)[123] と称する）がある．これはカメラを使わない（に限らない）一般の SLAM に関する研究で，移動体が未知環境を探索するとき，フィルタリングが行うように，過去の位置を周辺化せずそのまま変数として残しておいて，大域的な最適化を行なう．その際に前時刻における最適化の結果を更新することで，計算量を低減しようとする．写真測量/バンドル調整では，ちょうどこれと同じ方法があり[94]，それは，ヘッセ行列（情報行列）を新しい観測で更新した後，それをコレスキー分解する際に，以前の分解結果を利用する方法である．

また本節では扱わなかったが，Visual SLAM ので大規模空間を扱ったときの困難さを解決しようとした研究に，Pinies-Tardós[124]（単眼 SLAM）や，Konolige-Agarwal[125]（ステレオ SLAM）がある．さらに，ある場所を出発して未知空間を探索後，同じ場所に戻ってきたときに，そのことを検出する Loop-closing も，近年盛んに研究されている．その解決に，アピアランスを用いる方法が提案され[126),127)]，これは Appearance-only SLAM という，3 次元復元を行わない新しいジャンルの SLAM へとつながっている．

3. 多視点映像を用いた人物の全周囲３次元形状・運動復元

多視点映像には，物理的に複数のカメラを配置することで空間的に多視点を実現したものと，カメラを移動させながら静的なシーンを撮影することで時間的に多視点を実現したものの２種類が考えられるが，本節では特に前者の複数カメラ環境を用いて運動する人物の全周囲３次元形状・運動を復元する研究に注目する．

このような複数のカメラ映像から人物の時系列３次元形状を復元する研究は，対象のテクスチャ情報を用いるアプローチと，シルエット情報を用いるアプローチの２種類からスタートした．具体的には前者は通常のステレオ法[128),129)] をベースに，2.5 次元形状（depth-map）を貼り合わせることで全周囲３次元形状復元を行った Kanade らの研究[130] と 2.5 次元形状を介さずに直接３次元形状を求める Seitz らの研究（volumetric stereo）[131] が，後者は shape-from-silhouette[132),133),134)] で得られる visual hull をベースにした Moezzi らの研究[135] が，共に 90 年代後半に提案されている．

また 2000 年代に入るとテクスチャ情報とシルエット情報を同時に用いるアプローチが多く研究されるようになった[136),137),138),139),140),141),142),143),144),145),146),147)]．これはテクスチャマッチングによる形状復元が “視点間の対応付けが決まれば対象表面形状を正しく計算できるが，視点間の対応付けを常に正しく行うことは容易ではない” という特徴を有しているのに対して，シルエットを用いた形状復元は “対象の概形（visual hull）しか求まらないが，視点間の対応付けが不要で比較的安定に形状が求まる” という相補的な関係を持っているという分析に基づいている．すなわち，まず多視点シルエットから visual hull として “対象が必ず存在する範囲” を安定に求め，次いでこの範囲内でテクスチャの一致度（photo-consistency）を最大化する形状を求める，という考え方である．

また上記のように１時刻の３次元形状を独立に求めるだけでなく，複数時刻の形状を同時に復元する方法[148),149)] や，ある時刻における対象形状をまず復元し，これをキーフレームとして隣接する時刻の形状へと逐次変形することで形状と運動を同時に復元する方法[150),151)] のように３次元形状と同時に運動を推定する手法も研究されている．

以下では人物の全周囲３次元形状・運動復元を目的とした研究を，

入力　テクスチャ，シルエット
出力　１時刻の３次元形状のみ，隣接する２時刻の３次元形状，時系列３次元形状
形状表現　Voxel，メッシュ，パッチ集合
最適化　山登り法，coarse-to-fine，belief-propagation，graph-cuts, convex-optimization

© 2011 Information Processing Society of Japan

可視判定　State-based, oriented, 投票ベース
Photo-consistency　SSD, SAD, NCC, ZNCC
のような側面から分類してサーベイを行い，また最後に得られた時系列全周囲３次元形状を入力として用いた研究について述べる．

3.1 入　　力

3.1.1 テクスチャ

撮影画像のテクスチャ情報のみを用いた手法は，２次元画像平面上で画素を単位として対応付けを考える手法[130)152)]と，３次元空間中で voxel を単位として対応付けを考える手法[131)153)154)]の２つに分類することができる．

前者は従来のステレオ法からの自然な発展として各カメラ視点における 2.5 次元形状（depth-map）を計算し，これを３次元空間で貼り合わせることで３次元形状を得る手法である．この手法の特徴は非常に多くの研究がなされている depth-map 推定の成果[129)]をそのまま受け継ぐ点である．その一方，各視点で推定された depth-map の形状，および occlusion 判定の結果が，貼り合わせ（registration）の段階で互いに矛盾しないことが保証されない点に手法の困難さが存在する．また各カメラ視点で depth-map 推定を行うことができると仮定するため，比較的高い密度のカメラ配置が求められる（文献[130)]では直径 5m の半球上に 51 台）．

これに対して後者は対象が存在しうる空間を微小な単位体積（voxel）へと分割し，各 voxel が対象表面を構成するか否かをその voxel を各カメラ画像へと投影した際のテクスチャ一致度を用いて決定する手法であり space carving と呼ばれる．この定式化では前者のように貼り合わせというステップが不要となるため，各視点から見た形状の一貫性に関する問題は存在しない．ただし各 voxel のテクスチャ一致度を計算する際にどのカメラを使うのか，つまりその voxel を観測しているカメラをどのようにして推定するかという問題は未解決として残っている（後述）．

3.1.2 シルエット

次に入力としてシルエットを用いた手法は shape-from-silhouette (SfS) あるいは visual-cone-intersection と呼ばれ，その概念は繰り返し提案されてきた[132)133)134)]．この手法では各視点においてカメラの投影中心を頂点，撮影された対象シルエットを底面として持つ錐体（視錐体，visual cone）をまず定義し，これらの積領域を計算する．この積領域は視体積（visual hull）と呼ばれ，定義からその投影像は各視点のシルエットと一致する（入力シルエットに誤りが無かった場合）とともに，各視点のシルエットを投影像として与えうる最大の３次元形状である．そのためこの visual hull の計算をもって対象形状推定とするのが SfS の考え方である．

SfS の長所は (1) 視点間の対応付け処理を伴わない，(2) 静的な背景環境（特にブルーバックなど）ではシルエット抽出を安定に行うことができる，という理由から visual hull の計算が前述のテクスチャに基づくアプローチに比べて形状推定が頑健であると期待できる点にある．

一方で短所は，まずその定義から visual hull は各視点でシルエット輪郭として観測された形状のみを反映したものであり，シルエット輪郭としては観測できなかった部分（特に凹領域）は visual hull では反映されないことが挙げられる．この問題に対してはテクスチャ情報を併用する研究が多く提案されている（後述）．次に各視点のシルエットを投影像として与えうる最大の３次元形状として定義されるため，phantom volume と呼ばれる “偽領域” が生成される可能性がある．この偽領域も後述のテクスチャ情報を併用するアプローチで解決する可能性があるが，Miller らは “safe hull” と呼ばれる考え方によって phantom volume の一部を安全に除去する手法を提案している[155)156)]．また visual hull に形状に関する制約を入れることでより人体に近い３次元形状を得る手法[157)]も提案されている．この他にも積領域として計算することに起因する短所として，１つの視点でもシルエットに欠損（false-negative）があった場合には，該当する部分が visual hull から欠落するという問題がある（逆にシルエット中の false-positive 部分は，他の視点で true-negative であれば問題ない）．この問題に対して一度 visual hull を作成した後に欠損を修復する手法[158)]，シルエットと visual hull を同時に推定する手法[159)160)161)]などが提案されている．

SfS の実装は大きく (1) 画素を単位として２次元平面シルエット輪郭を離散化する手法と，(2) ３次元空間を voxel で離散化する手法の２つに分けることができる．

前者はは visual cone をシルエット輪郭を用いてピクセル精度で構築し，それらの間の積領域を陽に計算する手法である[162)163)164)165)156)]．このアプローチの場合，元のシルエット解像度そのものを反映した高精度な visual hull を生成することができる．これに対して後者は各 voxel を各視点に投影し，１視点でもシルエット外に投影されていたらその voxel を削除するという考え方である．この手法はアルゴリズムがシンプルで並列化が容易なため，PC クラスタを用いた実時間処理[166)167)]や GPU による実装[168)]に適している．しかしその一方で visual hull の空間解像度は３次元空間を離散化する解像度で決定され，空間全体を均一に離散化した場合のメモリコストは解像度の３乗に比例する．このため octree のような適応的な離散化によってメモリコストを抑えながらも前者と同程度の解像度の visual

© 2011 Information Processing Society of Japan

hull を計算する手法が提案されている[169]．

3.1.3 複数情報を用いた手法

これまでに述べたようにシルエットを用いた手法とテクスチャを用いた手法の特徴は互いに相補的な関係にある．そこでまず SfS によって対象の概形を決定し，これの形状をテクスチャマッチングによって高精度化するというアプローチが多く提案されている[136)137)138)139)140)141)142)143)144)145)146)147)]．これらの研究の多くは visual hull をまず初期値として使用するとともに後述のように可視判定の手がかりとして使用する．また最終的な出力３次元形状を各視点に投影すると，撮影されたシルエットと一致するという制約を使用している．ただし撮影シルエットとの一致という制約条件はしばしば “シルエットの輪郭と一致する” と表現され，輪郭内部が埋まっているかどうかは明示的に取り扱われないことが多い．これに対して Sinha らは一度形状復元を行った後にその投影像と撮影シルエットを比較して欠損部分を判定し，これを補うように再度形状復元することでシルエット内部についても一致させる反復的なアプローチを提案した[143]．一方 Cremers らは形状復元の最適化計算過程において３次元形状とシルエット内部との一致を明示的に表現できる手法を提案している[147]．また Guillemaut らは３次元形状と多視点シルエットを同時に推定する手法を提案している[170]．

シルエットとテクスチャ以外の情報として，Fua らは陰影情報を用いる手法を提案している[171]．また Tung らは各カメラ視点における SfM とカメラ間の wide-baseline stereo，および SfS を組み合わせた手法を提案している[172]．

3.2 出　　力

ここまでに述べた手法は全て１つの時刻における対象形状を復元対象とするものであったが，多視点映像から時系列３次元形状を復元する手法も複数提案されている．Vedula らは隣接する２つの時刻の voxel 空間の直積６次元空間で space carving を行う手法を提案した[148]．６次元空間中の１点は時刻間の voxel の組に対応するため，この空間での space carving は２時刻の３次元形状とその間の運動を同時に復元したことに対応している．Goldlucke らは時系列３次元形状を４次元時空間中のサーフェースとして推定する手法を提案した[149]．この手法では各時刻の visual hull を初期値として用いているが，出力形状が時間的に連続となるように制約を入れることで，ある１時刻のみの多視点画像だけでは観測できなかった対象形状についても妥当な推定を行うことができることを示している．またシルエットやテクスチャを用いて３次元形状と運動を同時に推定する手法[173)174)] も提案されている．

一方上記のように多視点映像から複数時刻の形状・運動を同時に求めるのではなく，ある１時刻の形状をまず形状復元し，それが他時刻のテクスチャ・シルエット情報と一致するように変形させる手法も提案されている[150)144)]．この手法ではメッシュ頂点の移動による変形という操作が形状復元と運動復元を同時に行っていることになる．

また Iwashita らは対象形状を voxel で表すと共に level-set によってその運動を実時間推定する方法を提案している[175]．

3.3 形 状 表 現

形状を voxel 集合として表現するアプローチ[141)142)] のメリットの１つとして初期形状からのトポロジー変更が容易であることが挙げられる．これは前述のように visual hull を初期値として使用し，さらに phantom volume によってそのトポロジーが真の形状のトポロジーとは異なる場合に特に意味を持つ．また対象の表面形状ではなくその内側まで復元するとした場合，形状復元とは各 voxel を対象/非対象（前景/背景）のいずれかに分類する問題であると表現したことになる．これは後述の graph-cuts による２値のラベル付け問題と見なせば大局解が求まるという別のメリットをもたらす．

一方で対象表面形状をメッシュモデルで表し，これを変形させることで形状復元を行うこともできる[136)138)139)176]．メッシュモデルで表面形状を表すメリットは，要素間の連結関係や面の向きが明示的にモデル化されているために形状に関する制約を導入しやすい点が挙げられる．しかしこの点は逆に，メッシュ変形を単純に頂点移動で表現した場合には初期形状のトポロジーと真の形状とが一致している必要があるというデメリットにも繋がっている．この問題に対して Varanasi らはメッシュトポロジーの変形を明示的にモデル化したアルゴリズムを提案している[177]．

これらとは別に，高い確信度をもって復元できる微小面から復元を開始し，テクスチャが存在しないか，繰り返しパターンなどの要因で復元に曖昧さが残る部分を徐々に決めていく手法も提案されている[178)179)]．このようなパッチ集合としての表現は前述の voxel，メッシュそれぞれの表現の中間に位置すると解釈することができる．

3.4 最　適　化

2000 年代前半の voxel carving，メッシュ変形などのアプローチではいわゆる山登り探索的な局所最適化を行う手法が多く見られていた[131)153)137)138)139)] が，その後は level-set を用いたもの[180)] や voxel 表現における graph-cuts による最適化[154)] およびメッシュモデルにおける belief-propagation による最適化[176]，coarse-to-fine 戦略を用いたもの[140)179)]，また convex optimization を用いたもの[181)147)182)] などが提案されている．特に graph-cuts を用いたアプローチはグラフ構成や埋め込む制約条件の違いによって様々なバリエーション

© 2011 Information Processing Society of Japan

が提案されている[141)142)183)184)143)185)144)146)172)]．

ただし graph-cuts, belief-propagation, convex-optimization などはいずれも大局最適解もしくはそれから一定の誤差範囲内の近似解を求めるためのアルゴリズムであるが，3次元形状をこれらの手法を通して推定した結果が真に最適な，つまり photo-consistency を最大化する3次元形状であるという保証は存在しない点には注意が必要である．これは次節で述べるように photo-consistency を正しく計算するための可視判定と3次元対象形状が鶏と卵の関係になっていることから，上述の最適化計算中は何らかの形で可視判定を固定し，その仮説の下で評価関数のの最適化を行っているためである．また最適化の結果として得られる形状が大局最適解でない場合，大局最適解からの誤差範囲が保証されていたとしても，その誤差の大小と形状の物理的な推定精度が直接的に関係するわけではない点も注意が必要である．

3.5 可視判定

空間中の各点におけるテクスチャの一致度（photo-consistency）を計算するためには，その点を観測することのできるカメラを知る必要がある．これを可視判定（visibility）と呼ぶ．可視判定はその点がカメラの視野内に入っているかだけでなく，対象自身による自己遮蔽についても考慮しなくてはならない．つまり可視判定を正しく行うには真の対象形状が既知でなくてはならず，したがって両者は鶏と卵の関係となっている．そのため何らかの仮説に基づいて可視判定を近似し，photo-consistency を計算できるようにしなくてはならない．

最初に提案された方法は何らかの方法，多くの場合は visual hull として対象形状を仮に求め，空間中の各点における visibility は，visual hull 表面上の最近傍点の visibility によって近似するという考え方である[136)138)141)]．これは visual hull が対象のおおよその形を良く捉えている場合には，自己遮蔽を良く近似できるために visibility 判定も正しく機能すると期待できる．しかし例えば visual hull に phantom volume が含まれ，かつそれが真の表面形状とそれを観測するカメラとの間に存在する場合，phantom volume が “自己遮蔽” をするために正しい visibility 判定ができないという問題が生じる．また visual hull と対象の真の形状が大幅に異なる場合，特に visual hull が凹領域を復元できていない場合には，visual hull 表面上の最近傍点の visibility によって近似できるという仮説が成り立たない．例えば空のお椀のような形状の visual hull は，お椀の中が満たされたような形状となる．このときお椀の内壁にとって最近傍の点は外壁側の点となってしまい，完全に異なったカメラ群を可視判定で与えてしまう可能性が高い．

次に Lempitsky らは oriented-visibility という考え方を提案した[183)]．これは単純に仮定した点の位置と面の向きのみによってカメラから観測可能かどうかを判定するというものであり，state-based visibility の逆の特性を持っていると言える．

また Furukawa らの微小パッチからスタートして3次元表面形状を成長させていくアプローチ[179)]は，初期段階は oriented-visibility，途中からは state-based visibility を採用していると考えることができる．

このような考え方はいずれにせよ対象の形状や位置関係に基づいた可視判定法であるが，Vogiatzis らはこれと全く異なる投票に基づく可視判定法を提案した[186)]．これは空間中の各点を，ある範囲の方向から十分に多くのカメラで撮影できていたならば，自己遮蔽を検出できなかったことに起因するテクスチャマッチングの誤りは外れ値として見なせる程度にしか発生しないという仮説に基づいている．この方法は特にカメラが密に配置できている場合に有効であると考えられる．

3.6 Photo-consistency

Photo-consistency とは可視判定によって決定されたカメラ間で計算されたテクスチャの一致度のことを意味し，その計算には通常のステレオと同様に SAD (Sum of Absolute Differences), SAD (Sum of Squared Differences), NCC (Normalized Cross Correlation), ZNCC (Zero-Mean NCC) などが用いられる．

またこのような一致度尺度とは別に，ステレオ法におけるテンプレートの大きさと射影変換による変形を考慮するか，つまり微小面を仮定してその法線方向に合わせてマッチングウィンドウを変形させるかどうかなども特にカメラ配置が疎である場合には重要となる．例えば state-based-visibility の場合は visibility を与えた visual hull 上の点における法線方向を，oriented-visibility では空間を離散化した単位体積の各面の位置と向きがそのまま用いられる．またパッチ集合を用いた手法[179)178)]では，仮定する面の法線方向も最適化対象として推定される．

3.7 人物の時系列全周囲3次元形状を用いた研究

ここまでで述べた研究はいずれにおいても3次元形状復元を行うことを目的としていたが，近年は時系列3次元形状が得られたと仮定した上で，各時刻の形状同士をマッチングすることで運動を求める手法[187)188)189)190)]やキーフレームの変形（メッシュトラッキング）を繰り返すことで運動を求める手法[177)191)]などが提案された．

またさらに骨格構造を埋め込むことで人物の多関節剛体としての運動を推定する研究も行われている[192)193)194)]

© 2011 Information Processing Society of Japan

参 考 文 献

1) Sameer Agarwal, Noah Snavely, Steven M. Seitz, and Richard Szeliski. Bundle adjustment in the large. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, Vol. 6312 of *Lecture Notes in Computer Science*, pp. 29–42. Springer, 2010.
2) S.Agarwal1, N.Snavely, I.Simon, S.Seitz, and R.Szeliski. Building Rome in a day. In *Proc. ICCV*, pp. 72–79, 2009.
3) A.Akbarzadeh, J.-M. Frahm, P.Mordohai, B.Clipp, C.Engels, D.Gallup, P.Merrell, M.Phelps, S.Sinha, B.Talton, L.Wang, Q.Yang, H.Stewénius, R.Yang, G.Welch, H.Towles, D.Nistér, and M.Pollefeys. Towards urban 3D reconstruction from video. In *Proc. 3DPVT*, May 2006.
4) Sunil Arya, David M. Mount, Nathan S. Netanyahu, Ruth Silverman, and Angela Y. Wu. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. In *ACM-SIAM SYMPOSIUM ON DISCRETE ALGORITHMS*, pp. 573–582, 1994.
5) A.Baumberg. Reliable feature matching across widely separated views. In *Proc. CVPR*, pp. 774–781, 2000.
6) H. Bay, A. Ess, T. Tuytelaars, and L.J. Van Gool. Speeded-up robust features (SURF). *CVIU*, Vol. 110, No.3, pp. 346–359, June 2008.
7) G.Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.
8) Martin Byröd and Kalle Åström. Conjugate gradient bundle adjustment. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, Vol. 6312 of *Lecture Notes in Computer Science*, pp. 114–127. Springer, 2010.
9) Anne-Laure Chauve, Patrick Labatut, and Jean-Philippe Pons. Robust piecewise-planar 3d reconstruction and completion from large-scale unstructured point data. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, Vol.0, pp. 1261–1268, 2010.
10) Jongmoo Choi and G.Medioni. Starsac: Stable random sample consensus for parameter estimation. *Proc. CVPR*, Vol.0, pp. 675–682, 2009.
11) St Q haneChristy, Radu Horaud. Euclidean reconstruction: from paraperspective to perspective. In B.Buxton and Roberto Cipolla, editors, *Proceedings of the 4th European Conference on Computer Vision, Cambridge, England*, pp. 129–140. Sprin-ger-Ver-lag, April 1996.
12) O.Chum and J.Matas. Matching with prosac: Progressive sample consensus. In *Proc. CVPR*, pp. I: 220–226, 2005.
13) Ondřej Chum and Jiří Matas. Optimal randomized ransac. *PAMI*, Vol.30, No.8, pp. 1472–1482, August 2008.
14) Ondřej Chum, Tomáš Werner, and Jiří Matas. Two-view geometry estimation unaffected by a dominant plane. In *Proc. CVPR*, Vol.1, pp. 772–779, 2005.
15) Yuchao Dai, Hongdong Li, and Mingyi He. Element-wise factorization for n-view projective reconstruction. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, Vol. 6314 of *Lecture Notes in Computer Science*, pp. 396–409. Springer, 2010.
16) M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM*, Vol.24, No.6, pp. 381–395, 1981.
17) Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, and Marc Pollefeys. Building rome on a cloudless day. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, Vol. 6314 of *Lecture Notes in Computer Science*, pp. 368–381. Springer, 2010.
18) Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.32, pp. 1362–1376, 2010.
19) Riccardo Gherardi, Michela Farenzena, and Andrea Fusiello. Improving the efficiency of hierarchical structure-and-motion. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, Vol.0, pp. 1594–1600, 2010.
20) Riccardo Gherardi and Andrea Fusiello. Practical autocalibration. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, Vol. 6311 of *Lecture Notes in Computer Science*, pp. 790–801. Springer, 2010.
21) Google: Image search. http://www.google.com/images, 2003.
22) C.G. Harris and M.Stephens. A combined corner and edge detector. In *Proc. Alvey Vision Conf.*, pp. 147–151, 1988.
23) R.I. Hartley and A.Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.
24) Richard I. Hartley. In defense of the eight-point algorithm. *PAMI*, Vol.19, pp. 580–593, 1997.
25) M.Havlena, A.Torii, J.Knopp, and T.Pajdla. Randomized structure from motion based on atomic 3D models from camera triplets. In *Proc. CVPR*, pp. 2874–2881, 2009.
26) M. Havlena, A. Torii, and T. Pajdla. Efficient structure from motion by graph optimization. In *Proc. ECCV*, 2010.
27) http://www.bing.com/.
28) http://www.flickr.com/.
29) Y.Ke and R.Sukthankar. PCA-SIFT: A more distinctive representation for local

© 2011 Information Processing Society of Japan

image descriptors. In *Proc. CVPR*, 2004.
30) E. Kruppa. Zur Ermittlung eines Objektes aus zwei Perspektiven mit innerer Orientierung. *Sitz.-Ber. Akad. Wiss., Wien, Math. Naturw. Abt. IIa*, Vol. 122, pp. 1939–1948, 1913.
31) Zuzana Kukelova, Martin Bujnak, and Tomas Pajdla. Automatic generator of minimal problem solvers. In *Proc. ECCV*, pp. 302–315, Berlin, Heidelberg, 2008. Springer-Verlag.
32) Zuzana Kukelova, Martina Bujnak, and Pajdla Tomas. Closed-form solutions to the minimal absolute pose problems with known vertical direction. In *Proc. ACCV*, 2010.
33) Hongdong Li. Multi-view structure computation without explicitly estimating motion. *Proc. CVPR*, Vol.0, pp. 2777–2784, 2010.
34) Hongdong Li and Richard Hartley. Five-point motion estimation made easy. In *Proc. ICPR*, ICPR '06, pp. 630–633, Washington, DC, USA, 2006. IEEE Computer Society.
35) X.Li, C.Wu, C.Zach, S.Lazebnik, and Frahm J.-M. Modeling and recognition of landmark image collections using iconic scene graphs. In *Proc. ECCV*, 2008.
36) H.C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, Vol. 293, pp. 133–135, 1981.
37) ManolisI.A. Lourakis. Sparse non-linear least squares optimization for geometric vision. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, Vol. 6312 of *Lecture Notes in Computer Science*, pp. 43–56. Springer, 2010.
38) M.I.A. Lourakis and A.A. Argyros. The design and implementation of a generic sparse bundle adjustment software package based on the Levenberg-Marquardt algorithm. Tech. Report 340, Institute of Computer Science – FORTH, August 2004.
39) D.G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, Vol.60, No.2, pp. 91–110, November 2004.
40) B.D. Lucas and T.Kanade. An iterative image registration technique with an application to stereo vision. In *Proc.of the 7th International Joint Conference on Artificial Intelligence*, pp. 674–679, 1981.
41) D.Martinec and T.Pajdla. Robust rotation and translation estimation in multiview reconstruction. In *Proc. CVPR*, 2007.
42) J.Matas, O.Chum, M.Urban, and T.Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proc. BMVC*, pp. 384–393, 2002.
43) Jiri Matas and Ondrej Chum. Randomized ransac with sequential probability ratio test. In *ICCV*, pp. 1727–1732, 2005.
44) Microsoft. Photosynth - http://livelabs.com/photosynth, 2008.
45) K.Mikolajczyk and C.Schmid. An affine invariant interest point detector. In *Proc. ECCV*. Springer-Verlag, 2002.
46) K.Mikolajczyk and C.Schmid. A performance evaluation of local descriptors. *IEEE PAMI*, 2004. submitted to PAMI.
47) K.Mikolajczyk, T.Tuytelaars, C.Schmid, A.Zisserman, J.Matas, F.Schaffalitzky, T.Kadir, and L.VanGool. A comparison of affine region detectors. *IJCV*, Vol.65, No. 1/2, pp. 43–72, 2005.
48) M.Muja and D.Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP*, 2009.
49) Kai Ni, Hailin Jin, and Frank Dellaert. Groupsac: Efficient consensus in the presence of groupings. In *Proc. ICCV*, pp. 2193–2200, 2009.
50) Kai Ni, Drew Steedly, and Frank Dellaert. Out-of-core bundle adjustment for large-scale 3d reconstruction. In *ICCV*, pp. 1–8, 2007.
51) D.Nistér. Preemptive RANSAC for live structure and motion estimation. In *Proc. ICCV*, pp. 199–206, 2003.
52) D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE PAMI*, Vol.26, No.6, pp. 756–770, 2004.
53) D.Nistér and F.Schaffalitzky. Four points in two or three calibrated views: theory and practice. *IJCV*, 2005. To appear.
54) D.Nister and H.Stewenius. Scalable recognition with a vocabulary tree. In *Proc. CVPR*, 2006.
55) Š.Obdržálek and J.Matas. Object recognition using local affine frames on distinguished regions. In *Proc. BMVC*, pp. 113–122, 2002.
56) A.Oliva and A.Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, Vol.42, No.3, pp. 145–175, May 2001.
57) ConradJ. Poelman and Takeo Kanade. A paraperspective factorization method for shape and motion recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.19, pp. 206–218, 1997.
58) L.Quan, B.Triggs, and B.Mourrain. Some results on minimal euclidean reconstruction from four points. *J. Math. Imaging and Vision*, 2003. To appear.
59) Rahul Raguram, Jan-Michael Frahm, and Marc Pollefeys. A comparative analysis of ransac techniques leading to adaptive real-time random sample consensus. In *Proc. ECCV*, pp. 500–513, Berlin, Heidelberg, 2008. Springer-Verlag.
60) Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *Proc. ECCV*, Vol.1, pp. 430–443, May 2006.
61) T.Sattler, B.Leibe, and L.Kobbelt. Scramsac: Improving ransac's efficiency with a spatial consistency filter. In *Proc. ICCV*, pp. 2090–2097, 2009.
62) F.Schaffalitzky and A.Zisserman. Multi-view matching for unordered image sets,

© 2011 Information Processing Society of Japan

or “How do I organize my holiday snaps?”. In *Proc. ECCV*, Vol.1, pp. 414–431. Springer-Verlag, 2002.
63) J.Sivic and A.Zisserman. Video Google: Efficient visual search of videos. In *CLOR*, pp. 127–144, 2006.
64) N.Snavely, S.Seitz, and R.Szeliski. Photo tourism: exploring photo collections in 3D. In *SIGGRAPH*, 2006.
65) N.Snavely, S.Seitz, and R.Szeliski. Modeling the world from internet photo collections. *IJCV*, Vol.80, No.2, pp. 189–210, 2008.
66) N. Snavely, S.M. Seitz, and R.S. Szeliski. Skeletal graphs for efficient structure from motion. In *Proc. CVPR*, 2008.
67) Noah Snavely. Bundler: Structure from motion (sfm) for unordered image collections. http://phototour.cs.washington.edu/bundler/, 2008.
68) Noah Snavely. *Scene Reconstruction and Visualization from Internet Photo Collections*. PhD thesis, University of Washington, 2008.
69) H.Stewénius, D.Nistér, F.Kahl, and F.Schaffalitzky. A minimal solution for relative pose with unknown focal length. In *Proc. CVPR*, 2005.
70) P.Sturm and W.Triggs. A factorization based algorithm for multi-image projective structure and motion. In *Proc. ECCV*, pp. 709–720, 1996.
71) Martin Bujnak Tomas Pajdla, Zuzana Kukelova. Minimal problems in computer vision.
72) C.Tomasi and T.Kanade. Shape and motion from image streams under orthography: A factorization approach. *IJCV*, Vol.9, No.2, pp. 137–154, 1992.
73) P.H.S. Torr. An assessment of information criteria for motion model selection. In *Proc. CVPR*, 1997.
74) P.H.S. Torr and A.Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *CVIU*, Vol.78, pp. 138–156, 2000.
75) W. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment: A modern synthesis. In W.Triggs, A.Zisserman, and R.Szeliski, editors, *Vision Algorithms: Theory and Practice*, LNCS, pp. 298–375. Springer Verlag, 2000.
76) A. Vedaldi and B. Fulkerson. Vlfeat – an open and portable library of computer vision algorithms. In *Proc. ACM international conference on Multimedia*, 2010.
77) Changchang Wu. SiftGPU: A GPU implementation of scale invariant feature transform (SIFT). http://cs.unc.edu/ccwu/siftgpu, 2007.
78) Christopher Zach, Manfred Klopschitz, and Marc Pollefeys. Disambiguating visual relations using loop constraints. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, Vol.0, pp. 1426–1433, 2010.
79) 岡谷貴之. コンピュータビジョン最先端ガイド 3 (CVIM チュートリアルシリーズ), 第 1 章. アドコム・メディア, 2010.
80) 植芝俊夫, 岡谷貴之, 佐藤智和. カメラキャリブレーション (サーベイ (1)). 情報処理学会研究報告. CVIM, [コンピュータビジョンとイメージメディア], Vol. 2005, No.18, pp. 1–18, 2005-03-03.
81) 藤吉弘亘. Gradient ベースの特徴抽出 - sift と hog -. Technical report, 情報処理学会 研究報告 CVIM 160, 2007.
82) Bartoli, A. and Strum, P.: Multi-view structure and motion from line correspondences, *Proc. ICCV* (2003).
83) Rosten, E. and Drummond, T.: Fusing Points and Lines for High Performance Tracking, *Proc. ICCV*, Springer, pp.1508–1515 (2005).
84) Chandraker, M., Lim, J. and Kriegman, D.: Moving in Stereo: Efficient Structure and Motion Using Lines, *Proc. ICCV* (2009).
85) Strasdat, H., Montiel, J. M.M. and Davison, J.: Real-time Monocular SLAM: Why Filter?, *Proc. ICRA* (2010).
86) Gennery, D.B.: Tracking known three-dimensional objects, *Proc. AAAI 2nd Natl. Conf. Artif. Intell.*, pp.13–17 (1982).
87) Gennery, D.B.: Visual tracking of known three dimensional objects, *Internaitonal Journal of Computer Vision*, Vol.7, No.3, pp.243–270 (1992).
88) Dickmanns, E. D. and Graefe, V.: Applications of dynamic monocular machine vision, *Machine Vision and Applications*, Vol.1, pp.1002–1029 (1988).
89) Harris, C. and Pike, J.M.: 3D positional integration from image sequences, *Proc. 3rd Alvey Vision Conf.*, pp.233–236 (1987).
90) Broida, T.J., Chandrashekhar, S. and Chellappa, R.: Recursive 3-D motion estimation from a monocular image sequence, *IEEE Transactions on Aerospace and Electronic Systems*, Vol.26, No.4, pp.639–656 (1990).
91) Azarbayejani, A. and Pentland, A.P.: Recursive Estimation of Motion, Structure, and Focal Length, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol.17, pp.562–575 (1995).
92) Chiuso, A. and Soatto, S.: MFm: 3-D Motion From 2-D Motion Causally Integrated Over Time, *Proc. ECCV*, pp.735–750 (2000).
93) McLauchlan, P.F. and Murray, D.W.: A unifying framework for structure and motion recovery from image sequences, *Proc. ICCV*, pp.314– (1995).
94) Triggs, B., McLauchlan, P., Hartley, R. and Fitzgibbon, A.: Bundle Adjustment — A Modern Synthesis, *Vision Algorithm: Theory & Practice* (Triggs, B., Zisserman, A. and Szeliski, R., eds.), Springer-Verlag LNCS 1883 (2000).
95) Mclauchlan, P.F.: A Batch/Recursive Algorithm for 3D Scene Reconstruction, *Proc. CVPR*, pp.738–743 (2000).
96) Davison, A.J.: Real-Time Simultaneous Localisation and Mapping with a Single Camera, *Proc. ICCV*, pp.1403– (2003).

© 2011 Information Processing Society of Japan

97) Davison, A.J., Reid, I.D., Molton, N.D. and Stasse, O.: MonoSLAM: Real-Time Single Camera SLAM, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.29, No.6, pp.1052–1067 (2007).
98) Davison, A.J. and Kita, N.: 3D simultaneous localisation and map-building using active vision for a robot moving on undulating terrain, *Proc. CVPR* (2001).
99) Shi, J. and Tomasi, C.: Good features to track, *Proc. CVPR*, pp.539–600 (1994).
100) Civera, J., Davison, A.J. and Montiel, J. M.M.: Inverse Depth Parametrization for Monocular SLAM., *IEEE Transactions on Robotics*, Vol.24, No.5, pp.932–945 (2008).
101) Maybank, S.J.: Filter Based Estimates of Depth, *Proc. BMVC*, pp.349–354 (1990).
102) Bailey, T., Nieto, J., Guivant, J.E., Stevens, M. and Nebot, E.M.: Consistency of the EKF-SLAM Algorithm., *Proc. IROS*, IEEE, pp.3562–3568 (2006).
103) Montemerlo, M.: FastSLAM: A Factored Solution to the Simultaneous Localization and Mapping Problem with Unknown Data Association, PhD Thesis, Robotics Institute, Carnegie Mellon University (2003).
104) Eade, E. and Drummond, T.: Scalable Monocular SLAM, *Proc. CVPR*, pp.469–476 (2006).
105) Bailey, T., Nieto, J. and Nebot, E.: Consistency of the FastSLAM algorithm, *Proc. ICRA*, pp.424–429 (2006).
106) Eade, E. and Drummond, T.: Monocular SLAM as a Graph of Coalesced Observations, *Proc. ICCV*, pp.1–8 (2007).
107) Rosten, E. and Drummond, T.: Machine learning for high-speed corner detection, *Proc. ECCV*, pp.430–443 (2006).
108) Fitzgibbon, A.W. and Zisserman, A.: Automatic Camera Recovery for Closed or Open Image Sequences, *Proc. ECCV*, pp.311–326 (1998).
109) Nistér, D.: Reconstruction from uncalibrated sequences with a hierarchy of trifocal tensors, *Proc. ECCV*, pp.649–663 (2000).
110) Torr, P. H.S., Fitzgibbon, A.W. and Zisserman, A.: The problem of degeneracy in structure and motion recovery from uncalibrated image sequences, *International Journal of Computer Vision*, Vol.32, No.1, pp.27–44 (1999).
111) Shum, H.-Y., Zhang, Z. and Ke, Q.: Efficient Bundle Adjustment with Virtual Key Frames: A Hierarchical Approach to Multi-Frame Structure from Motion, *Proc. CVPR*, p.2538 (1999).
112) Zhang, Z., Zhang, Z., Shan, Y. and Shan, Y.: Incremental Motion Estimation Through Local Bundle Adjustment, *Proc. ICIP* (2003).
113) Engels, C., Stewénius, H. and Nistér, D.: Bundle adjustment rules, *Proc. Photogrammetric Computer Vision* (2006).
114) Mouragnon, E., Lhuillier, M., Dhome, M., Dekeyser, F. and Sayd, P.: Real Time Localization and 3D Reconstruction, *Proc. CVPR*, pp.363–370 (2006).
115) Mouragnon, E., Lhuillier, M., Dhome, M., Dekeyser, F. and Sayd, P.: Generic and real-time structure from motion using local bundle adjustment, *Image and Vision Computing*, Vol.27, No.8, pp.1178–1193 (2009).
116) Klein, G. and Murray, D.: Parallel Tracking and Mapping for Small AR Workspaces, *Proc. ISMAR* (2007).
117) Nistér, D., Naroditsky, O. and Bergen, J.: Visual Odometry, *Proc. CVPR*, pp. 652–659 (2004).
118) Nistér, D.: An Efficient Solution to the Five-Point Relative Pose Problem, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol.26, pp.756–777 (2004).
119) Oliensis, J.: Exact two-image structure from motion, *IEEE Trans. PAMI*, Vol.24, No.12, pp.1618–1633 (2002).
120) Haralick, B. M., Lee, C.-N., Ottenberg, K. and Nölle, M.: Review and analysis of solutions of the three point perspective pose estimation problem, *International Journal of Computer Vision*, Vol.13, No.3, pp.331–356 (1994).
121) Klein, G. and Murray, D.: Improving the Agility of Keyframe-Based SLAM, *Proc. ECCV*, pp.802–815 (2008).
122) Klein, G. and Murray, D.: Parallel Tracking and Mapping on a Camera Phone, *Proc. ISMAR* (2009).
123) Kaess, M., Ranganathan, A. and Dellaert, F.: iSAM: Incremental Smoothing and Mapping, *IEEE Trans. on Robotics*, Vol.24, No.6, pp.1365–1378 (2008).
124) Pinies, P. and Tardós, J.D.: Large scale SLAM building conditionally independent local maps: Application to monocular vision, *IEEE Trans. Robotics*, Vol.24, No.5, pp.1094–110 6 (2008).
125) Konolige, K. and Agrawal, M.: FrameSLAM: From Bundle Adjustment to Real-Time Visual Mapping, *IEEE Transactions on Robotics*, Vol.24, No.5, pp.1066–1077 (2008).
126) Cummins, M. and Newman, P.: Probabilistic appearance based navigation and loop closing, *Proc. ICRA* (2007).
127) Angeli, A., Filliat, D., Doncieux, S. and Meyer, J.A.: Fast and incremental method for loop-closure detection using bags of visual words, *IEEE Trans. Robotics*, Vol.24, No.5, pp.1027–1037 (2008).
128) Okutomi, M. and Kanade, T.: A locally adaptive window for signal matching, *IJCV*, Vol.7, pp.143–162 (1992).
129) Scharstein, D. and Szeliski, R.: A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms, *IJCV*, Vol.47, pp.7–42 (2002).
130) Kanade, T., Rander, P. and Narayanan, P.J.: Virtualized Reality: Constructing Virtual Worlds from Real Scenes, *IEEE Multimedia*, pp.34–47 (1997).

© 2011 Information Processing Society of Japan

131) Seitz, S. and Dyer, C.: Photorealistic Scene Reconstruction by Voxel Coloring, *IJCV*, Vol.25, No.3, pp.151–173 (1999).
132) Baumgart, B.G.: A polyhedron representation for computer vision, *Proceedings of the National Computer Conference and Exposition*, AFIPS '75, pp.589–596 (1975).
133) Martin, W.N. and Aggarwal, J.K.: Volumetric description of objects from multiple views, *PAMI*, Vol.5(2), pp.150–158 (1983).
134) Laurentini, A.: The visual hull concept for silhouette-based image understanding, *PAMI*, Vol.16, No.2, pp.150 –162 (1994).
135) Moezzi, S., Tai, L.-C. and Gerard, P.: Virtual View Generation for 3D Digital Video, *IEEE Multimedia*, pp.18–26 (1997).
136) Isidoro, J. and Sclaroff, S.: Stochastic Mesh-Based Multiview Reconstruction, *Proc. of 3DPVT*, Padova, Italy, pp.568–577 (2002).
137) Cheung, K.M., Baker, S. and Kanade, T.: Visual Hull Alignment and Refinement Across Time: A 3D Reconstruction Algorithm Combining Shape-From-Silhouette with Stereo, *Proc.of CVPR*, pp.375–382 (2003).
138) Matsuyama, T., Wu, X., Takai, T. and Nobuhara, S.: Real-Time 3D Shape Reconstruction, Dynamic 3D Mesh Deformation and High Fidelity Visualization for 3D Video, *CVIU*, Vol.96, pp.393–434 (2004).
139) Esteban, C.H. and Schmitt, F.: Silhouette and stereo fusion for 3D object modeling, *CVIU*, Vol.96, pp.367–392 (2004).
140) Sinha, S.N. and Pollefeys, M.: Multi-View Reconstruction Using Photo-consistency and Exact Silhouette Constraints: A Maximum-Flow Formulation, *Proc.of ICCV*, pp.349–356 (2005).
141) Starck, J., Hilton, A. and Miller, G.: Volumetric stereo with silhouette and feature constraints, *Proc.of BMVC*, pp.1189–1198 (2006).
142) Tran, S. and Davis, L.: 3D Surface Reconstruction Using Graph Cuts with Surface Constraints, *Proc.of ECCV*, Vol.3952, pp.219–231 (2006).
143) Sinha, S., Mordohai, P. and Pollefeys, M.: Multi-View Stereo via Graph Cuts on the Dual of an Adaptive Tetrahedral Mesh, *Proc.of ICCV*, pp.1 –8 (2007).
144) Furukawa, Y. and Ponce, J.: Carved Visual Hulls for Image-Based Modeling, *IJCV*, Vol.81, pp.53–67 (2009).
145) Starck, J., Maki, A., Nobuhara, S., Hilton, A. and Matsuyama, T.: The Multiple-Camera 3-D Production Studio, *IEEE Tran. on Circuit and Systems for Video Technology*, Vol.19, No.6, pp.856 –869 (2009).
146) Hisatomi, K., Tomiyama, K., Katayama, M. and Iwadate, Y.: Method of 3D reconstruction using graph cuts, and its application to preserving intangible cultural heritage, *IEEE Workshop on eHeritage and Digital Art Preservation*, pp.923 –930 (2009).
147) Cremers, D. and Kolev, K.: Multiview Stereo and Silhouette Consistency via Convex Functionals over Convex Domains, *PAMI*, No.99, p.1 (2010).
148) Vedula, S., Baker, S., Seitz, S. and Kanade, T.: Shape and Motion Carving in 6D, *Proc.of CVPR* (2000).
149) Goldlüecke, B. and Magnor, M.: Space-Time Isosurface Evolution for Temporally Coherent 3D Reconstruction, *Proc.of CVPR*, Washington, D.C., USA, pp.350–355 (2004).
150) Nobuhara, S. and Matsuyama, T.: Heterogeneous Deformation Model for 3D Shape and Motion Recovery from Multi-Viewpoint Images, *Proc.of 3DPVT*, Thessaloniki, Greece, pp.566–573 (2004).
151) Furukawa, Y. and Ponce, J.: Dense 3D motion capture from synchronized video streams, *Proc.of CVPR*, pp.1 –8 (2008).
152) Li, J., Li, E., Chen, Y., Xu, L. and Zhang, Y.: Bundled depth-map merging for multi-view stereo, *Proc.of CVPR*, Vol.0, pp.2769–2776 (2010).
153) Kutulakos, K.N. and Seitz, S.M.: A theory of shape by space carving, *Proc. of ICCV*, pp.307–314 (1999).
154) Vogiatzis, G., Torr, P. H.S. and Cipolla, R.: Multi-View Stereo via Volumetric Graph-Cuts, *CVPR*, pp.391–398 (2005).
155) Miller, G. and Hilton, A.: Safe Hulls, *Proc. 4th European Conference on Visual Media Production*, IET (2007).
156) Chen, G., Su, H., Jiang, J. and Wu, W.: Safe Polyhedral Visual Hulls, *Advances in Multimedia Modeling* (Boll, S., Tian, Q., Zhang, L., Zhang, Z. and Chen, Y.-P., eds.), Lecture Notes in Computer Science, Vol.5916, Springer Berlin / Heidelberg, pp.35–44 (2010).
157) Franco, J.-S., Lapierre, M. and Boyer, E.: Visual Shapes of Silhouette Sets, *Proc. of 3DPVT*, pp.397 –404 (2006).
158) Toyoura, M., Iiyama, M., Kakusho, K. and Minoh, M.: Silhouette Extraction with Random Pattern Backgrounds for the Volume Intersection Method, *Proc.of 3DIM*, pp.225–232 (2007).
159) Franco, J.-S. and Boyer, E.: Fusion of multiview silhouette cues using a space occupancy grid, *Proc.of ICCV*, Vol.2, pp.1747 –1753 Vol. 2 (2005).
160) Nobuhara, S., Tsuda, Y., Ohama, I. and Matsuyama, T.: Multi-viewpoint Silhouette Extraction with 3D Context-aware Error Detection, Correction, and Shadow Suppression, *IPSJ Transactions on Computer Vision and Applications*, Vol.1, pp. 242–259 (2009).
161) Campbell, N., Vogiatzis, G., C.Hern 疣 dez, Cipolla, R. ∶ Automatic 3D object segmentation in multiple views using volumetric graph-cuts, *Image and Vision Computing*, Vol.28, No.1, pp.14 – 25 (2010).

© 2011 Information Processing Society of Japan

162) Matusik, W., Buehler, C. and McMillan, L.: Polyhedral Visual Hulls for Real-Time Rendering, *Proceedings of the 12th Eurographics Workshop on Rendering Techniques*, pp.115–126 (2001).
163) Boyer, E. and Franco, J.-S.: A hybrid approach for computing visual hulls of complex objects, *Proc.of CVPR*, Vol.1, pp.I–695 – I–701 vol.1 (2003).
164) Lazebnik, S., Furukawa, Y. and Ponce, J.: Projective Visual Hulls, *IJCV*, Vol.74, pp.137–165 (2007).
165) Franco, J.-S. and Boyer, E.: Efficient Polyhedral Modeling from Silhouettes, *PAMI*, Vol.31, No.3, pp.414 –427 (2009).
166) Cheung, K.M., Kanade, T., Bouguet, J.-Y. and Holler, M.: A Real Time System for Robust 3D Voxel Reconstruction of Human Motions, *Proc. of CVPR*, South Carolina, USA, pp.714–720 (2000).
167) Matsuyama, T., Wu, X., Takai, T. and Wada, T.: Real-Time Dynamic 3D Object Shape Reconstruction and High-Fidelity Texture Mapping for 3D Video, *IEEE Tran. on Circuit and Systems for Video Technology*, Vol.14, pp.357–369 (2004).
168) Li, M., Magnor, M. and Seidel, H.-P.: Hardware-Accelerated Visual Hull Reconstruction and Rendering, *Proceedings of Graphics Interface (GI'03)* (2003).
169) Liang, C. and Wong, K.-Y.K.: Exact Visual Hull From Marching Cubes, *Proc.of VISAPP*, pp.597–604 (2008).
170) Guillemaut, J.-Y., Kilner, J. and Hilton, A.: Robust graph-cut scene segmentation and reconstruction for free-viewpoint video of complex dynamic scenes, *Proc. of ICCV*, pp.809 –816 (2009).
171) Fua, P. and Leclerc, Y.G.: Using 3-Dimensional Meshes To Combine Image-Based and Geometry-Based Constraints, *Proc.of ECCV*, pp.281–291 (1994).
172) Tung, T., Nobuhara, S. and Matsuyama, T.: Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo, *Proc.of ICCV*, pp.1709 –1716 (2009).
173) Guan, L., Franco, J.-S., Boyer, E. and Pollefeys, M.: Probabilistic 3D occupancy flow with latent silhouette cues, *Proc.of CVPR*, pp.1379 –1386 (2010).
174) Pons, J.-P., Keriven, R. and Faugeras, O.: Multi-View Stereo Reconstruction and Scene Flow Estimation with a Global Image-Based Matching Score, *IJCV*, Vol.72, pp.179–193 (2007).
175) Iwashita, Y., Kurazume, R., Hara, K., Uchida, S., Morooka, K. and Hasegawa, T.: Fast 3D reconstruction of human shape and motion tracking by parallel fast level set method, *Proc.of ICRA*, pp.980 –986 (2008).
176) Vogiatzis, G., Torr, P., Seitz, S.M. and Cipolla, R.: Reconstructing Relief Surfaces, *Proc.of BMVC*, pp.117–126 (2004).
177) Varanasi, K., Zaharescu, A., Boyer, E. and Horaud, R.: Temporal Surface Tracking Using Mesh Evolution, *Proc.of ECCV*, Vol.5303, pp.30–43 (2008).
178) Habbecke, M. and Kobbelt, L.: A Surface-Growing Approach to Multi-View Stereo Reconstruction, *Proc.of CVPR*, Vol.0, pp.1–8 (2007).
179) Furukawa, Y. and Ponce, J.: Accurate, Dense, and Robust Multi-View Stereopsis, *Proc.of CVPR*, pp.1–8 (2007).
180) Soatto, S., Yezzi, A.J. and Jin, H.: Tales of Shape and Radiance in Multi-view Stereo, *Proc.of ICCV*, pp.974–981 (2003).
181) Kolev, K., Klodt, M., Brox, T. and Cremers, D.: Continuous Global Optimization in Multiview 3D Reconstruction, *International Journal of Computer Vision*, Vol.84, No.1, pp.80–96 (2009).
182) Kolev, K., Pock, T. and Cremers, D.: Anisotropic Minimal Surfaces Integrating Photoconsistency and Normal Information for Multiview Stereo, *Proc. of ECCV* (2010).
183) Lempitsky, V., Boykov, Y., Ivanov, D. and Ivanov, D.: Oriented Visibility for Multiview Reconstruction, *Proc.of ECCV*, pp.226–238 (2006).
184) Boykov, Y. and Lempitsky, V.: From Photohulls to Photoflux Optimization, *Proc. of BMVC*, p.III:1149 (2006).
185) Hernandez, C., Vogiatzis, G. and Cipolla, R.: Probabilistic visibility for multi-view stereo, *Proc.of CVPR*, pp.1 –8 (2007).
186) Vogiatzis, G., Hernandez, C., Torr, P. and Cipolla, R.: Multiview Stereo via Volumetric Graph-Cuts and Occlusion Robust Photo-Consistency, *PAMI*, Vol.29, No.12, pp.2241 –2246 (2007).
187) Starck, J. and Hilton, A.: Correspondence labelling for wide-timeframe free-form surface matching, *Proc.of ICCV*, pp.1 –8 (2007).
188) Mateus, D., Horaud, R., Knossow, D., Cuzzolin, F. and Boyer, E.: Articulated shape matching using Laplacian eigenfunctions and unsupervised point registration, *Proc.of CVPR*, pp.1 –8 (2008).
189) Zaharescu, A., Boyer, E., Varanasi, K. and Horaud, R.: Surface feature detection and description with applications to mesh matching, *Proc.of CVPR*, pp.373 –380 (2009).
190) Tung, T. and Matsuyama, T.: Dynamic surface matching by geodesic mapping for 3D animation transfer, *Proc.of CVPR*, pp.1402–1409 (2010).
191) Cagniart, C., Boyer, E. and Ilic, S.: Free-form mesh tracking: A patch-based approach, *Proc.of CVPR*, pp.1339 –1346 (2010).
192) Menier, C., Boyer, E. and Raffin, B.: 3D Skeleton-Based Body Pose Recovery, *Proc.of 3DPVT*, pp.389 –396 (2006).
193) Mukasa, T., Miyamoto, A., Nobuhara, S., Maki, A. and Matsuyama, T.: Complex human motion estimation using visibility, *Proc.of FG*, pp.1 –6 (2008).

© 2011 Information Processing Society of Japan

Vol.2011-CVIM-176 No.1
2011/3/17

194) Horaud, R., Niskanen, M., Dewaele, G. and Boyer, E.: Human Motion Tracking by Registering an Articulated Surface to 3D Points and Normals, *PAMI*, Vol.31, No.1, pp.158 –163 (2009).

© 2011 Information Processing Society of Japan