解説

解説：Structure from Motion (SfM)

第一回　SfM の概要とバンドル調整

織田　和夫*

SfM という用語を学会誌や学術講演会で頻繁に耳にするようになった。SfM 系ソフトウェアで画像処理を行うことによって点群やオルソフォトが自動的に簡単に得られるようになったことは写真測量のすそ野を広げることになった。その一方で SfM と写真測量が全く異なる技術であるかのような誤解があることも事実である。本解説は 3 回にわたって SfM について解説し，SfM を利用する上での基本的な知識を提供する。

全 3 回の予定は以下のとおりである。

①SfM の概要とバンドル調整

SfM の概要と，その大きな構成要素であるバンドル調整について概観する。

②SfM と多視点マッチング

MVS を例に SfM における多視点マッチング技術について概説する。

③SfM 系写真処理ソフトウェア

代表的な SfM 系写真処理ソフトウェアとその機能的特徴について概説する。

1　SfM 概論

1.1　SfM とは

画像を取得すると，そこには 3 次元空間を反映する様々な情報が含まれる。画像に映った対象の 3 次元的形状を画像から得る方法として，影を用いる方法（Shape from Shading）や，ピントを利用する用いる方法（Shape from Defocus）などがあり，これらをまとめたものが Shape from X という概念である。移動（Motion）するカメラから得られる画像から形状を復元するのが SfM（Shape from Motion）である。Structure from Motion（SfM）は Shape from Motion の別名である。ただし，Structure from Motion というと，画像に映った対象物の幾何学形状とカメラの動きを同時に復元する手法という意味合いが強くなる[1]。以下では SfM は Structure from Motion の略称とする。

SfM はもともとコンピュータビジョンやロボットビジョンからきた概念である。カメラの位置姿勢と対象の座標を取得するということであれば，写真測量における空中三角測量と SfM との違いはないように思えるかもしれない。ただし SfM は本質的にはセンサやコンピュータを用いて外界の情報を把握することにある。明確に定義されているわけではないが，基本的に SfM は“自動的”に処理を行うことを前提にしている。SfM はコンピュータやロボットの視覚として利用できるよう，極力処理が自動化されていることが本質的である。

ロボットビジョンでは周囲の 3 次元構造と自分の位置を推定する技術のことを SLAM（Simultaneous Localization and Mapping）と呼んでいる。SLAM にはレーザセンサを利用したものもあるが，画像を使用するもの（Visual SLAM）[2]の計測原理は SfM と同じといってよい。

1.2　SfM ソフトウェアの処理過程

SfM ソフトウェアの典型的な処理過程を図 1 に示す。カメラの位置・姿勢推定とタイポイントの 3 次元

*アジア航測株式会社

座標の算出の際中心となるのはバンドル調整である。すなわち，写真測量の空中三角測量（Aerial Triangulation）と同様である。ソフトウェアによってはマニュアルで基準点位置の入力やタイポイントの入力を行うこともできるが，基本的にはすべての処理が自動的に行われる。SfM というとここまでの処理を指すことも多い。

カメラの位置・姿勢推定が終わると，多視点画像計測による点群計測（Dense Stereo Matching）が行われる。3 枚以上の画像も同時に利用しながら自動マッチング（多視点画像計測）を行うため，2 枚の画像を用いた自動マッチングよりもオクルージョン（建物等で地表面上に見えない部分が生じること）の影響を受けにくい。結果として 3 次元色付点群が得られる。

ソフトウェアによっては更に自由表面形状モデリングを行い，テクスチャ付ポリゴンモデルとして出力する。SfM 系ソフトウェアで出力したテクスチャ付ポリゴンモデルの例を図 2 に示す。

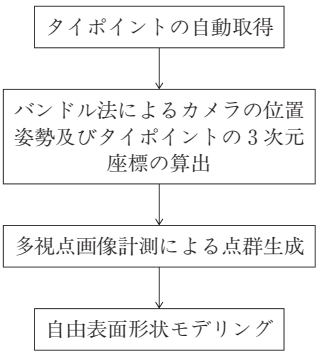

図 1　SfM の処理フロー

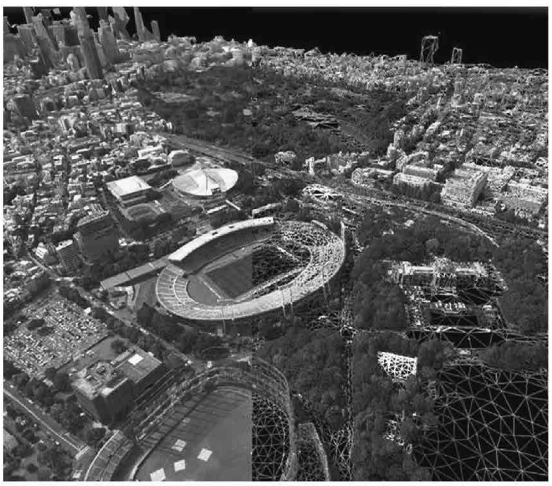

図 2　斜め撮影航空写真カメラによる 3 次元モデル

SfM では処理を自動的に行うが，画像の撮影条件によっては自動処理が失敗する場合がある。例えばオーバラップ・サイドラップが少ない場合や，画像内にテクスチャが極端に少ない場合，大きな水域が含まれている場合などである。自動処理が失敗するとそれより先に処理を進めることができないので注意が必要である。

2　SfM とバンドル調整

2.1　動画像を用いた SfM

SfM は 1 台のカメラが動きながら撮影するというところから研究が始まっている。そのため当初は動画像を利用した研究が中心であった。

SfM を行う上で重要なテクニックは，一連の画像の間で自動的に対応点（同じ場所が移っている点）を抽出することである。動画像の場合，特徴点の移動が少ないので，離散的に撮影された画像よりはるかに対応点抽出が行いやすい。連続的に撮影された画像（動画像）から特徴点の移動を見出すにはオプティカルフローが利用される[3]。

動画像を用いた SfM の代表的な手法としてあげられるのは，Tomasi-Kanade の因子分解法[1]（Factorization Method）である。因子分解法は，一連の特徴点（i=1～M）が一連の画像（j=1～N）上で (x_{ij}, y_{ij}) として観測されたとき，この (x_{ij}, y_{ij}) で構成される行列

$$P=\begin{pmatrix} x_{11} & \cdots & x_{1M} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{NM} \\ y_{11} & \cdots & y_{1M} \\ \vdots & \ddots & \vdots \\ y_{N1} & \cdots & y_{NM} \end{pmatrix} \tag{1}$$

を，画像の投影を決める行列 M と 3 次元位置情報を表す行列 S に因子分解する。

$$P=M\cdot S \tag{2}$$

因子分解法は連続的に点を画像上で追跡することによって処理できること，また線代数的処理によって簡便に実装できるという利点がある。

2.2　静止画像を利用した SfM

因子分解法が研究されていた1990年台前半は，民生

品のデジタルカメラが普及する前であり，対象画像もSDサイズの小さい画像であった1990年台後半に民生品デジタルカメラの普及がすすみ，またカメラが携帯電話にも搭載されるようになると同時に，インターネットが爆発的に普及するようになると，多くのデジタルカメラ画像が利用対象として取り上げられるようになった。形状の計測においてはなるべく異なった視点から画像を取得し，画像間で対応点を取る方が精度的に有効であり，使用する画像の有効画素数が多い方がより精細な解析を行うことができるという面でもデジタルカメラで得られる静止画像は動画像より有利であるといえる。

1990年台半ばから爆発的に普及が進んだインターネットには，デジタル画像が大量にアップロードされるようになった。これらの画像はSfMの研究の上で大きな役割を果たした。パブリックドメインの代表的なソフトウェアであるBundlerは，ワシントン州立大学とマイクロソフト社で行われたPhoto Toolismというプロジェクトで開発された[4]。これは著名な観光資源（たとえばローマのコロッセアム）を撮影した画像をインターネット上から検索・収集し，これらから3次元モデルを生成するというものである。

デジタルカメラ画像は連続的に撮影されるわけではないので，対応点探索でオプティカルフローを用いることはできない。異なる場所・異なる向き・異なるスケールで撮影された画像間で自動的に対応点を取ることができる技術として開発されたのがSIFT[5]などに代表される画像局所特徴量を利用した対応点探索である。

これらの画像局所特徴量は画像内の特徴点の特徴をパラメータ化したもので，回転やある程度のスケール変化に影響を受けずに特徴点間の類似性を算出することが可能である。また処理対象の画像は同一のカメラで取得されている必要はなく，異なるカメラで撮影した画像も取り扱うことが可能になった。当初のSfMで想定されていた移動するカメラだけでなく，異なる時期・異なるカメラで撮影した様々な画像も取り扱う対象となった。このことがSfMの有用性を大きく拡張することとなった。

オプティカルフローで求めた対応点に比べて自動的に求められた対応点には誤った対応点も大量に含まれている。対応点を抽出するだけでなく，それから正しい対応点を抽出することが重要である。正しい対応点を抽出するのに利用されるのがRANSAC[6]（Random Sample Consensus）である。RANSACは無作為に取り出したサンプルに満たすべき拘束条件を適用し，拘束条件を満たすそれ以外のサンプルがなるべく多くなるようにする。これによって拘束条件（共線条件もしくはエピポーラ拘束）を満たさなかった点を排除することができる。

たとえばRANSACを2枚の画像の対応点に適用するとする。任意の8点を選ぶと，それから8点法を用いて基礎行列（Fundamental Matrix）を求めることができる。基礎行列$\mathbf{H}$は，2枚の画像1上の対応点(x_1', y_1')と(x_2', y_2')の間のエピポーラ拘束を表現する3×3の行列である。

$$(x_1' y_1', 1)\cdot\mathbf{H}\cdot\begin{pmatrix} x_2' \\ y_2' \\ 1 \end{pmatrix} = (x_1', y_1', 1)\cdot\begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix}\cdot\begin{pmatrix} x_2' \\ y_2' \\ 1 \end{pmatrix} = 0 \tag{3}$$

もとめた基礎行列を選択した8点以外の対応点に適用した際に上記のエピポーラ拘束を満たす（実際にはある閾値以内で満たす）対応点が指定された試行回数内で最大になるとき，この際にエピポーラ拘束を満たさない対応点は誤対応として取り除く。

2.3 バンドル調整とSfM

現在流通しているSfMを実装したソフトウェアは，撮影された画像の位置と姿勢，および画像間の対応点の3次元座標の計算を行うのにあたってバンドル調整を採用している。

バンドル調整は言うまでもなく写真測量における空中三角測量で用いられている手法である。バンドル調整は，カメラの内部標定要素も含め非常にフレキシブルかつ精密な調整が可能である。ただし，非線形最小二乗法で最適化を行うための外部標定要素（カメラ位置・姿勢）および内部標定要素（カメラパラメータ）の初期値をある程度正しく与える必要がある。

デジタルカメラの普及は，SfMの世界でバンドル調整を普及する上で（特に内部標定要素を与える上で）大きな影響を与えたと考えられる。JFIF形式（拡張されたJPEG形式）で保存された画像には，デジタルカメラで撮影した際の諸元（カメラ機種名・撮影時の焦点距離等）がおさめられている。カメラの機種名がわ

かれば，撮影素子のサイズの情報を得ることができる。よって，画像座標を，画像中心からの実寸座標であらわし，また概略の画面距離も与えることができる。これによって，画像座標から（概略の）写真座標への変換が可能となる。

外部標定要素については，SfM の世界では基本的に任意座標で計算する。外部標定要素の初期値は，例えば次のような手順で与えることができる[7]。

⑴ 相互標定を最初のペアで計算する。
⑵ バンドル調整を行う。
⑶ ⑵で得られた 3 次元座標を用いて DLT によってその他の画像の外部標定要素を計算する。
⑷ 与えられたすべての画像について処理できるまで⑵および⑶を繰り返す。

ここで最初のペアの選定と相互標定の実行が一つの問題となる。なぜなら相互標定は写真測量においては非線形最小二乗法を用いて行うため，収斂撮影のような場合に安定的に計算することができないからである。SfM において，相互標定を代数的に安定的に計算できる手法として 5 点法[8]（five point algorithm）が開発されている。

最初のペアを選ぶときに，同じ視点からの画像が選定されると相互標定および 3 次元座標の計算が不安定となる。これを避けるため，対応点同士が射影変換（homography）で変換されないようなペアを選定すればよい。

なお，UAV など屋外で撮影した画像については，GPS 等で得られた撮影位置が記録されており，商用の SfM 系ソフトウェアではこれらを利用することができる。

2.4 内部標定要素の取り扱い

SfM 系ソフトウェアで注意しなければならないのは内部標定要素の取り扱いである。写真測量においては，レンズディストーションは観測された写真座標の関数として扱われる。対して，SfM 系ソフトウェアでは，レンズディストーションなしで投影された位置の関数で与えられることが多い。例えば Bundler においては，共線条件から得られる写真座標 $\mathbf{p}$ と観測された座標 $\mathbf{p}'$ との間の関係は画面距離 f を用いての式で与えられる[9]。

$$\mathbf{p}' = f \cdot r(\mathbf{p}) \cdot \mathbf{p} \tag{4}$$

ここで $r(\mathbf{p})$ はレンズディストーションを与えるスケール関数であり，ベクトルの長さ $||\mathbf{p}||$ を用いて

$$r(\mathbf{p}) = 1.0 + k_1 \cdot ||\mathbf{p}||^2 + k_2 \cdot ||\mathbf{p}||^4 \tag{5}$$

である。

（受付日2016.5.9，受理日2016.6.17）

参考文献

1）Tomasi, C. and Kanade, T. (1992) : Shape and motion from image streams under orthography : a factorization method, Int'l. J. Computer Vision, 9(2), pp.137-154.

2）Iketani, A., Sato, T., Ikeda, S., Kanbara, M., Nakajima, N., and Yokoya, N. (2007) : Video mosaicing based on structure from motion for distortion-freedocument digitization, Proc. Asian Conf. on Computer Vision (ACCV2007), Vol. II, pp.73-84.

3）Lucas, B.D. and Kanade, T. (1981) : An Iterative Image Registration Technique with an Application to Stereo Vision, in Proc. of Int. Joint Conf. on Artificial Intelligence, pp.674-679.

4）Snavely, N., Seitz, S.M., and Szeliski, R. (2006) : Photo Tourism : Exploring image collections in 3D, ACM Transactions on Graphics (Proceedings of SIGGRAPH 2006).

5）Lowe, D. (2004) : Distinctive image features from scaleinvariant keypoints, Int'l. J. Computer Vision, 60, 2, pp.91-110.

6）Fischler, M. and Bolles, R. (1987) : Random sample consensus : a paradigm for model fitting with application to image analysis and automated cartography, Readings in computer vision : issues, problems, promciples, and paradigms, pp. 726-740.

7）Snavely, N., Seitz, S.M., and Szeliski, R. (2007) : Modeling the World from Internet Photo Collections. Int'l. J. Computer Vision.

8）Nister, D. (2004) : An efficient solution to the five-point relative pose problem, PAMI, 26(6), pp. 756-770.

9）http://www.cs.cornell.edu/~snavely/bundler/