Optical Engineering

OpticalEngineering.SPIEDigitalLibrary.org

Assessing geoaccuracy of structure from motion point clouds from long-range image collections

David Nilosek
Derek J. Walvoord
Carl Salvaggio

SPIE.

Optical Engineering 53(11), 113112 (November 2014)

Assessing geoaccuracy of structure from motion point clouds from long-range image collections

David Nilosek,[a,*] Derek J. Walvoord,[b] and Carl Salvaggio[a]
[a]Rochester Institute of Technology, 54 Lomb Memorial Drive, Rochester, New York 14623, United States
[b]Exelis Inc., 400 Initiative Drive, Rochester, New York 14606-0488, United States

Abstract. Automatically extracted and accurate scene structure generated from airborne platforms is a goal of many applications in the photogrammetry, remote sensing, and computer vision fields. This structure has traditionally been extracted automatically through the structure-from-motion (SfM) workflows. Although this process is very powerful, the analysis of error in accuracy can prove difficult. Our work presents a method of analyzing the georegistration error from SfM derived point clouds that have been transformed to a fixed Earth-based coordinate system. The error analysis is performed using synthetic airborne imagery which provides absolute truth for the ray-surface intersection of every pixel in every image. Three methods of georegistration are assessed; (1) using global positioning system (GPS) camera centers, (2) using pose information directly from on-board navigational instrumentation, and (3) using a recently developed method that utilizes the forward projection function and SfM-derived camera pose estimates. It was found that the georegistration derived from GPS camera centers and the direct use of pose information from on-board navigational instruments is very sensitive to noise from both the SfM process and instrumentation. The georegistration transform computed using the forward projection function and the derived pose estimates prove to be far more robust to these errors.
 [DOI: 10.1117/1.OE.53.11.113112]

Keywords: georegistration; point clouds; error analysis; structure from motion; photogrammetry; computer vision.

Paper 140832P received May 23, 2014; accepted for publication Oct. 21, 2014; published online Nov. 27, 2014.

1 Introduction

Identifying objects within a scene is a key goal in the field of computer vision. One method of describing objects within a scene is to identify their structure through analysis of their motion between multiple images, a process commonly referred to as structure from motion (SfM). This technique of analyzing objects has its roots in traditional photogrammetry, though the standard goal of SfM techniques lies in object identification rather than mensuration. To that end, the SfM algorithm chain assumes little or no information about the imaging platform. The computer vision community has developed a number of complex processes to estimate camera pose, i.e., the sensor position and orientation, but these methods are limited to estimating parameters in a relative sense.[1,2] Consequently, any estimated object structure is in the same relativistic coordinate system. Precise geographic measurements of objects, often required in photogrammetric applications, cannot be directly extracted in this coordinate system without additional information.

In many practical SfM systems, it is assumed that the camera's calibration information is known; this includes the focal length, pixel pitch, and sensor size.[1,3] Knowledge of this information allows for a metric reconstruction of both the camera pose and object structure.[1] In other words, ambiguity in absolute position, orientation, and scale are present in this process, and geographic information is necessary to register the metric coordinate system to a fixed, Earth-based coordinate system. Fortunately, the relationship between these coordinate systems can be described by a simple, seven degrees-of-freedom (DoF) similarity transform, $\mathbf{T}_s$, as they are both metric coordinate systems.

Although methods for SfM derived point cloud georegistration are known, researchers have started to question the accuracy of these methods. A number of researchers used accurate ground truth measurements for accuracy assessment. Neitzel and Klonowski[4] generated a dense collection of survey points, identified the corresponding points in the SfM georegistered reconstruction, and used that to generate error residuals. Crandall et al.[5] took a similar approach using highly accurate camera pose information as ground truth for comparison. Hudzietz and Saripalli[6] used known distance measurements along specific paths. Ground truth can also be collected in the form of range scanning. Koutsoudis et al.[7] took this approach in evaluating commercial SfM-based software, using a single structure as a test case.

Using ground truth as a means for accuracy assessment is a valid approach to this problem. However, this approach cannot isolate the effects of specific sources of error in the SfM process, e.g., noise in the recorded pose information versus error in feature matching. This work proposes a new technique for assessing the accuracy of an SfM-derived structure using a completely synthetic dataset. This dataset is generated using a first principles physics-based modeling engine alongside a highly detailed spectrally attributed three-dimensional (3-D) model. A synthetic dataset allows for complete control over all camera pose information, including the noise. Furthermore, the modeling engine collects ground truth for every pixel in every generated image, which can be used for error analysis.

*Address all correspondence to: David Nilosek, E-mail: drn2369@rit.edu

0091-3286/2014/$25.00

1.1 *Georegistration Methods*

With innovations in global positioning system (GPS) technology, images are often geotagged, and this additional information can be exploited in different ways to obtain a geographically accurate structure. Many SfM applications use imagery, openly available on the Internet, with geotags that have been manually recorded or automatically captured by the GPS of the device itself. Differences between close-range and long-range collection geometry play a large role in how this information is used in georegistration.

The simplest georegistration method uses just GPS information to extract $\mathbf{T_s}$ by matching the estimated camera pose to its corresponding GPS location.[8] Measurement uncertainties and outliers from most GPS devices drive the use of noise-reducing model optimization algorithms such as random sample consensus (RANSAC).[9] The error in this estimation alone is often too large, and additional information is used to further refine the transform. Wang et al.[10] incorporate Google Street View imagery and Google Earth models to further refine the position of an SfM-derived structure. Use of accurate digital surface models can also reduce positional error in ground-based SfM point clouds, as shown by Wendel et al.[11] It is even possible to incorporate the geolocations into the SfM process itself. Crandall et al.[5] utilize Internet-based imagery with geolocations to aid in the estimation of camera pose and orientation.[5] They propose an SfM technique that solves the problem using markov random fields by incorporating the geotags into pose and orientation estimation using an energy function that minimizes the noise present in these types of geotags.

Much of the research related to georegistration of SfM-derived point clouds focuses on imagery collected through devices which may have unreliable position and orientation information. Furthermore, many of these methods assume that distortion effects must be estimated. Imagery collected for photogrammetric use often comes from highly calibrated imaging systems on platforms flown with GPS devices and state-of-the-art inertial navigation systems (INS), which provides high-fidelity estimates of the absolute position and orientation. Having the ability to put significantly more trust into a sensor model derived from accurate information allows for a simplified georegistration approach to be taken, as presented by Walvoord et al.[12]

This work compares three methods of georegistration using the novel analysis technique: Simply triangulating using GPS/INS information, estimating $\mathbf{T_s}$ using camera centers,[8] and estimating $\mathbf{T_s}$ using a combination of both.[12] For the relevant discussion of georegistered SfM point clouds, a review of the SfM problem and current solutions, as well the estimation process for $\mathbf{T_s}$ is presented in the following sections.

2 Structure from Motion Problem

The goal in SfM is to design a robust, unsupervised scene reconstruction methodology from a collection of images. The SfM process can be described as a chain of individual processes which are formed together to perform the desired task, as shown in Fig. 1. A standard SfM processing chain immediately seeks to establish a geometric relationship between image pairs by applying a feature extraction algorithm to each image. The most common is the scale-invariant feature transform (SIFT) algorithm[13] due to its ability to provide a robust feature descriptor across image conditions. The next step in the SfM chain is to match the descriptors from each image to each other image, effectively computing correspondences between images. There are a variety of techniques available, from brute-force feature matching, to model-fitting algorithms that employ RANSAC.[9]

At this point, image-to-image correspondences permit estimation of a series of fundamental matrices, which provide the necessary epipolar geometry for cursory triangulation. A coarse estimate of the 3-D sparse point cloud provides a series of equations that relate the image coordinate system to the world coordinate system (WCS). The remainder of the problem lies in refining camera projection matrices to ensure consistent triangulation and accurate relative orientation. This critical step is accomplished using a form of bundle adjustment, a nonlinear optimization process typically used for selfcalibration and parameter refinement problems in projection functions.[14]

Using this approach of estimating the fundamental matrix, performing triangulation, and iteratively refining the solution allows the scene to be reconstructed up to a projective ambiguity. In other words, the scene reconstruction is determined at best to within a projective transformation with respect to the WCS. In the field of computer vision, it is well known that methods exist[1] to refine or "upgrade" the reconstruction to a metric reconstruction, in which the scene is determined up to a rotation, translation, and uniform scaling. Some SfM chains require knowledge of the intrinsic camera parameters prior to adjustment to achieve metric reconstruction.

It is worth noting that using position and orientation estimates from a GPS/INS device to seed the initial parameter vector will likely fail to produce a high-fidelity, geoaccurate point cloud. This may seem counterintuitive, as these estimates are close approximations to the true camera pose in the desired fixed, Earth-based coordinate system. However, forcing the camera pose to be consistent with GPS/INS measurements typically leads to parameter adjustments that produce greater inconsistencies between image correspondence, which leads to poor triangulation. Careful selection of the weights applied to adjustable parameters may help to reduce this error, but it is important to recognize that an initial estimate of the camera pose using correspondences to perform a resection guides the subsequent refinement toward a solution that, loosely speaking, agrees with the image-based geometry. The errors caused by GPS/INS measurements are explored in further detail within this work.

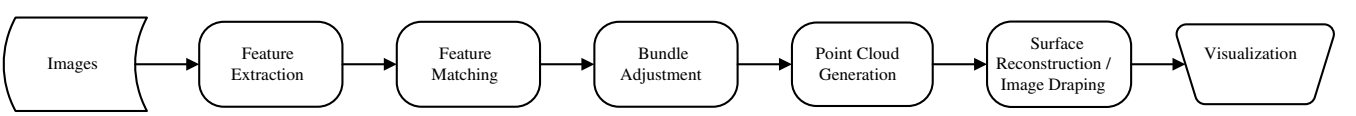


Fig. 1 High-level diagram of a procedural chain used to perform the structure-from-motion (SfM) task.

2.1 *SfM Software*

This work makes use of a specific open source SfM software workflow commonly used for dense point cloud generation. The workflow is composed of three programs, Bundler,[3] clustering views for multiview stereo (CMVS),[15] and patch-based multiview stereo (PMVS).[16] The methodology presented in this paper for accuracy analysis is applicable to any structure generation algorithm, however, the output from this specific workflow is used in the examples presented here.

2.1.1 *Bundler*

Bundler, one of the most commonly used open source SfM software packages, is a well-established SfM algorithm written by Noah Snavley.[3] Bundler uses an iterative bundle adjustment approach to perform the optimization. The input to the bundle adjustment is generated using initial image correspondences from SIFT, and the position and orientation of each view are then estimated using the five-point camera pose estimation algorithm.[17] Using this estimate, each image correspondence is triangulated with each image using a linear triangulation estimation. The correspondence, camera pose, and triangulated points are used as inputs to the bundle adjustment, which provides an error-minimized solution for each input. This is done initially for the two cameras with the most correspondences and is repeated by adding the camera with the next highest number of correspondences. By iteratively removing outliers, the parameter estimation process avoids the heavy influence of data that would otherwise disturb a least squares cost function in the bundle adjustment. The result of the Bundler process is an error-minimized set of camera pose information for each image, as well as a sparse point cloud generated from the initial SIFT correspondences.

2.1.2 *PMVS and CMVS*

Bundler provides a sparse reconstruction of the scene. However, a dense reconstruction is often desired, and PMVS software attempts to solve this problem.[16,15] This algorithm uses the camera pose provided by Bundler to narrow down a correspondence search, which allows for a large number of pixels to be matched and reconstructed. PMVS uses a patch-based model to estimate surface orientation and this model is also used as a measure of point correspondence. It has two majors steps; the first is an initial estimation of surface through sparse feature matching, guided by the camera pose information. The second step is a region growing process where the initial surface is grown and guided by the camera pose information to produce a dense scene reconstruction.

The PMVS algorithm is a multicore implementation which allows for the fast simultaneous processing of many images. However, when the image set grows large, this processing time can still become very long. A clustering algorithm called CMVS is implemented to break down the image set into manageable clusters.[15] CMVS is a graph-cut–based clustering algorithm which clusters the cameras according to their pose.[18] The goal is to find clusters of cameras which are observing the same region of the scene, and run PMVS on just the corresponding images. This not only reduces runtime for PMVS, but also the error in the resulting point cloud. The analysis tests performed in Sec. 5 use the Bundler, PMVS, and CMVS workflow to generate structure from imagery.

3 Error Analysis using Synthetic Data

The ability to control every parameter within an image collection is necessary to fully understand how georegistration methods perform. To this end, a synthetic aerial image dataset was generated using the digital imaging and remote sensing image generation (DIRSIG) model. This model was created by researchers from the digital imaging and remote sensing laboratory at the Rochester Institute of Technology.[19] DIRSIG is a first principles synthetic image generation model which can produce imagery ranging from the visible to the thermal infrared spectrum. The synthetic image generation process requires a spectrally attributed 3-D model of the scene to be imaged. This work utilized a variant of Megascene 1, a hand-created 3-D scene of a small suburban region in northeast Rochester, New York.[20] Larger block-style apartment buildings were added to the scene to increase the depth of targets within the scene.[21]

Many photogrammetric applications deal with only nadir-looking imagery, therefore, analysis of this type of imagery is explored in this work. A small 12 image dataset was created which contained 80% forward overlap and 60% side overlap. The sensor flying height and pixel pitch were designed such that the images had a ground sample distance of 0.3 m. These parameters were chosen to be representative of a densely flown nadir-looking collect.

DIRSIG uses a ray tracer to calculate the sensor response at each pixel in every synthetic frame, then the pixel to ground intersections for each ray is recorded and can be used as ground truth. Each synthetic frame is explicitly defined with calibration parameters and an exterior orientation, which is used to create a noiseless sensor model. An example image from the data used in this work along with an SfM derived point cloud is shown in Fig. 2.

Every 3-D point in the SfM process is created from corresponding pixels in multiple images. These pixels are tracked and used with the DIRSIG truth imagery to calculate the error in each 3-D point. The average of every truth measurement is used for error calculation. The Euclidean distance between the georegistered 3-D point and its corresponding truth point is taken as the error. The absolute truth values for every pixel, along with a noiseless camera pose, allow for an in-depth analysis of error within geoaccurate SfM processes.

4 Methods of Generating Geoaccurate Point Clouds

The application of the analysis described in Sec. 3 is demonstrated in this work using three different methods of point cloud georegistration. Two of these methods are based on estimating the similarity transform, $\mathbf{T_s}$, whereas the third method uses only the recorded GPS/INS information.

4.1 *Similarity Transform Estimation*

Georegistration of point clouds through a similarity transform to obtain absolute orientation has existed for many years and is well documented.[22,23] Points contained within two different Euclidean coordinate systems can be related using a similarity transform. A set of points $\mathbf{X}_A$ in a metric

Fig. 2 Digital imaging and remote sensing image generation (DIRSIG) is used to generate synthetic imagery which is input to the SfM process. Panels (a)–(c) are three close up renderings of Megascene, the three-dimensional (3-D) model used to generate the synthetic imagery. (d) A sample DIRSIG image created for this work along with the corresponding Z-dimension truth. (e) A view of the SfM derived point cloud.

coordinate system A, can be transformed to the desired metric coordinate system D using a seven DoF transform as shown

$$\mathbf{X}_D = s\mathbf{R}\mathbf{X}_A + \mathbf{t}, \tag{1}$$

where s is a uniform scale, $\mathbf{R}$ is a rotation matrix, and $\mathbf{t}$ is a translation vector. This transform can be calculated by using a set of known corresponding points, $\mathbf{x}_a$ and $\mathbf{x}_d$, such that $\mathbf{x}_a \in \mathbf{X}_D$ and $\mathbf{x}_d \in \mathbf{X}_D$. The first step in solving for this transform is calculating the scale. This is done by moving the centroid of each set of points, $\mathbf{x}_a$ and $\mathbf{x}_d$, to the origin of their respective coordinate systems, defined in Eqs. (2) and (3)

$$\mathbf{C}_{a,i} = \mathbf{x}_{a,i} - \bar{\mathbf{x}}_a, \tag{2}$$

$$\mathbf{C}_{d,i} = \mathbf{x}_{d,i} - \bar{\mathbf{x}}_d. \tag{3}$$

The scale value can then be determined by the ratio between the mean lengths of each of the zero-centered points set. Figure 3 shows an example of $\mathbf{C}_{a,i}$ and $\mathbf{C}_{d,i}$.

The scale relating the two point sets, $\mathbf{C}_{a,i}$ and $\mathbf{C}_{d,i}$, can be calculated through the ratio of the mean vectors lengths for each point set, as shown in Eq. (4) and pictured in Fig. 3:[24]

$$s^2 = \frac{\sum_{i=1}^{n} \left\| C_{d,i} \right\|^2}{\sum_{i=1}^{n} \left\| C_{a,i} \right\|^2}. \tag{4}$$

The next step is to compute the rotation matrix $\mathbf{R}$ which relates $\mathbf{C}_{a,i}$ to $\mathbf{C}_{d,i}$. This can be done using the Kabsch algorithm, which is a method for calculating an optimal rotation matrix between two sets of points in a least squares sense.[25,26] After the rotation matrix and uniform scale value are calculated the translation vector can be calculated using the centroids, as shown in Eq. (5):

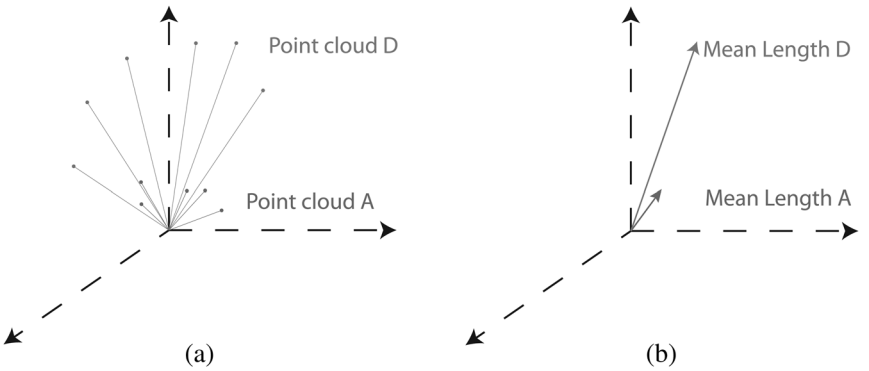


Fig. 3 (a) An example of two sets of zero centered point clouds. (b) The mean vector length of each set of points. The ratio of the mean lengths can be used to calculate the scale between the two data sets.

$$t = \bar{\mathbf{x}}_a - s\mathbf{R}\bar{\mathbf{x}}_d. \tag{5}$$

This method alone is sensitive to error in the corresponding points $\mathbf{x}_a$ and $\mathbf{x}_d$. The similarity transform can calculated using a model fitting method robust to outliers such as RANSAC.[9]

Extracting a subset of corresponding points from sets $\mathbf{X}_A$ and $\mathbf{X}_D$ is necessary for the calculation of a similarity transform. For the purposes of obtaining a georegistered coordinate system with a similarity transform, $\mathbf{X}_A$ is the SfM derived WCS and $\mathbf{X}_D$ is the desired fixed Earth-based coordinate system. The remainder of this section will discuss two methods of finding these subsets which utilize additional information known about the scene and its cameras. A third method using direct triangulation is also described.

4.2 *Using Camera Position Estimates*

The method most commonly used in the computer vision community uses geo-tags from imagery.[27] Estimated camera centers derived from the SfM process are used as the points from set $\mathbf{X}_A$, and the GPS-located camera centers are used as the points from set $\mathbf{X}_D$. The transformation process, which is henceforth referred to as the camera centers approach, is shown in Fig. 4.

Although this process requires only knowledge of the GPS information from each camera, it is highly susceptible to error. Given error in the GPS location or error in the camera pose estimation, the resulting transform is susceptible to significant inaccuracies. These errors can be mitigated using an outlier removal algorithm such as RANSAC.[5]

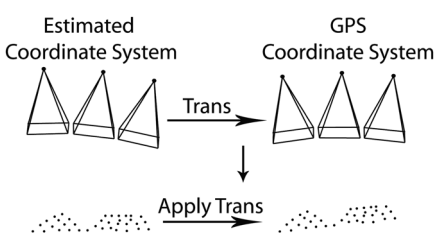


Fig. 4 The most common method for obtaining corresponding points utilizes the camera centers. Estimated 3-D coordinates from the SfM process and global positioning system (GPS)-located coordinates are used to calculate the transform. This transform is then used to convert the SfM estimated structure into the GPS coordinate system.

4.3 *Using the Forward Projection Function and Camera Pose Estimates*

The algorithm discussed hereafter, referred to as the augmented camera model approach, assumes an application that differs from a majority of SfM problems in both collection geometry and available information.[12] In addition to the sensor position information used by the algorithm in Sec. 4.2, it is assumed that the imagery has been captured using a platform with a GPS/INS device, and position and orientation information are readily available alongside a known forward projection function.[12] The initial point cloud is obtained by applying the SfM methodology described in Sec. 2, with the result having been iteratively refined to ensure consistency with the image-based geometry.

It should be noted that even perfect image correspondences will fail to triangulate if the camera position and orientation do not agree with some fixed geometry, in this case, the image-based geometry [Fig. 5(b)]. The SfM approach forces image features to lie on corresponding epipolar planes; this is commonly referred to as "optimal triangulation." Any attempt to control the camera position or orientation (with metadata) will modify this image-based geometry. Rather, the SfM reconstruction may now be placed in an Earth-based coordinate system using a simple two step process.

First, a low-fidelity sparse point cloud is generated to serve as a reference in the desired fixed Earth-based coordinate system, $\mathbf{x}_d$, where the geometry of this process is shown in Fig. 5(b). Three pieces of critical information enable this multi-image triangulation:

1. Refined pixel correspondences across multiple views from the SfM workflow
2. Image metadata containing GPS/INS position and orientation information
3. Complete knowledge of the ground-to-image function of the collection system

A simple direct linear triangulation algorithm[1] is easily extensible to correspondences across multiple views. New camera projection matrices corresponding to each view are derived from the physical sensor model and the available metadata. To avoid numerical instability due to matrices with a poor condition number, a normalization matrix is formulated for each view that effectively centers pixel

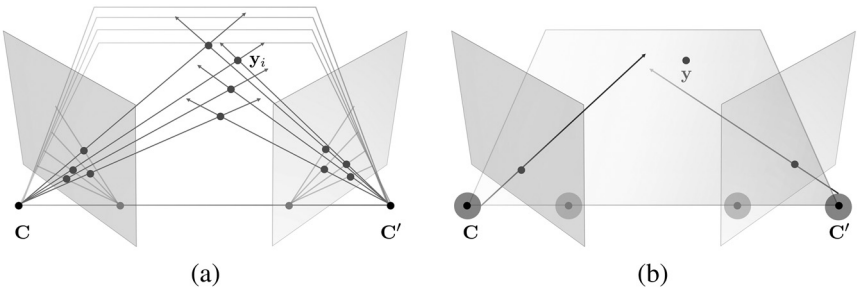


Fig. 5 (a) Triangulation using refined image-based geometry. Rays extending from **C** and **C**′ through image feature points consistent with the refined image-based geometry intersect at a 3-D point $\mathbf{y}_i$ in the epipolar plane. This is illustrated for several two-ray triangulations, which produces a point cloud in an arbitrary WCS. (b) Triangulation in the presence of errors in camera position and orientation. Rays extending from **C** and **C**′ through image feature points on corresponding epipolar lines do not intersect at a point in the epipolar plane due to discrepancies between the metadata and image-based geometry.

measurements (from that view) and scales the mean magnitude to $\sqrt{2}$. Each two-dimensional image point $\mathbf{u}_{i,j}$ may be expressed as a mapping from a 3-D point $\mathbf{x}_i \in \mathbf{X}_D$ in the fixed Earth-based coordinate system through the 3×4 camera projection matrix $\mathbf{P}_j$ for a particular view j. In homogeneous coordinates

$$\mathbf{u}_{i,j} = \mathbf{P}_j \mathbf{x}_i, \tag{6}$$

which can be rewritten as a cross product

$$\mathbf{0} = \mathbf{u}_{i,j} \otimes \mathbf{P}_j \mathbf{x}_i. \tag{7}$$

Using the previous equation, the system of equations above may expressed as[1]

$$\begin{bmatrix} u_{i,1}\mathbf{p}_{3,1}^{\mathrm{T}}\mathbf{x}_i - \mathbf{p}_{2,1}^{\mathrm{T}}\mathbf{x}_i \\ v_{i,1}\mathbf{p}_{3,1}^{\mathrm{T}}\mathbf{x}_i - \mathbf{p}_{1,1}^{\mathrm{T}}\mathbf{x}_i \\ u_{i,2}\mathbf{p}_{3,2}^{\mathrm{T}}\mathbf{x}_i - \mathbf{p}_{2,2}^{\mathrm{T}}\mathbf{x}_i \\ v_{i,2}\mathbf{p}_{3,2}^{\mathrm{T}}\mathbf{x}_i - \mathbf{p}_{1,2}^{\mathrm{T}}\mathbf{x}_i \\ \vdots \\ u_{i,n}\mathbf{p}_{3,n}^{\mathrm{T}}\mathbf{x}_i - \mathbf{p}_{2,n}^{\mathrm{T}}\mathbf{x}_i \\ v_{i,n}\mathbf{p}_{3,n}^{\mathrm{T}}\mathbf{x}_i - \mathbf{p}_{1,n}^{\mathrm{T}}\mathbf{x}_i \end{bmatrix} = \mathbf{0} \tag{8}$$

or simply

$$\mathbf{A}\mathbf{x}_i = \mathbf{0}. \tag{9}$$

The vector $\mathbf{x}_i$ that minimizes $||\mathbf{A}\mathbf{x}||$ subject to the condition $||\mathbf{x}|| = 1$ is the unit eigenvector with the smallest eigenvalue of the matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, i.e., the last column of $\mathbf{V}$ in the singular value decomposition $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}$; this is the i'th triangulated point. The geometry of this process is shown in Fig. 6(a). The process is repeated for all suitable image feature correspondences established in the SfM framework. Care should be taken to ensure that the assumed coordinate system of the focal plane array is consistent (or accounted for) between the SfM approach, e.g., Bundler,[27] and the physical sensor model. Two point clouds now exist: one high-fidelity point cloud in an arbitrary WCS, $\mathbf{x}_a$, and one low-fidelity point cloud in a fixed Earth-based coordinate system, $\mathbf{x}_d$. There is a one-to-one mapping between each 3-D point, as seen in Fig. 6(b).

The second step is to determine the transformation between these two coordinate systems. Ideally, the two point clouds are related by a translation, a uniform scale factor, and a rotation, but there is uncertainty present in both data sets. Image feature correspondences used for scene reconstruction have an associated error value from the final error vector ε of the L-M bundle adjustment solution after convergence. This error vector ε may be used to select a desired number of correspondences with the lowest image-based triangulation error, effectively reducing the size of each point cloud (and computation time of the similarity transform). The transformation that maps the relative

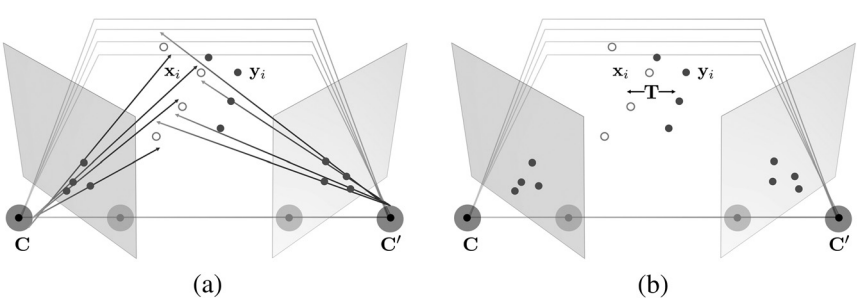


Fig. 6 (a) Triangulation using the physical sensor model and metadata. Rays extending from the camera centers, with uncertainty in position and orientation, through image feature points consistent with the refined image-based geometry intersect at a 3-D point $\mathbf{x}_i$ in the epipolar plane. This is illustrated for several two-ray triangulations, which produces a point cloud in a fixed Earth-based coordinate system. (b) Point cloud transformation. Points from both triangulation methods are related by a similarity transformation matrix **T**. This matrix maps the point cloud $\mathbf{y}_i$ to the coordinate system of the point cloud $\mathbf{x}_i$.

point cloud to the triangulated point clouds in the fixed Earth-based coordinate system is computed using the method presented in Sec. 4.1.

4.4 *Direct Triangulation Using Navigation Data*

Perhaps the simplest method of generating georegistered point clouds is to triangulate corresponding image points through the sensor model of the camera or cameras used to capture the image collection if navigation data is available. In many applications, particularly airborne collections, a GPS/INS device provides estimates of the position and orientation information at the time of capture. However, even in well-calibrated models, the random error present in these measurements leads to some degree of uncertainty in the triangulated result (see Fig. 5).

5 Analysis of Georegistration Techniques

Each method for calculating a georegistration transform is analyzed using the synthetic data analysis as described in Sec. 3. Bundler along with PMVS, as described in Sec. 2.1, is used as the SfM engine to generate a dense 3-D point cloud for georegistration. In addition, each image correspondence is triangulated using the camera pose information as described in Sec. 4.4 and is also analyzed. Finally, the georegistration error for each method is calculated again after random noise was added to the recorded camera pose data in order to simulate a more realistic scenario. The amount of random noise was determined based on RMS error reported in the Applanix POSTrack performance summary.[28]

5.1 *Georegistration Error using DIRSIG Noiseless Sensor Model*

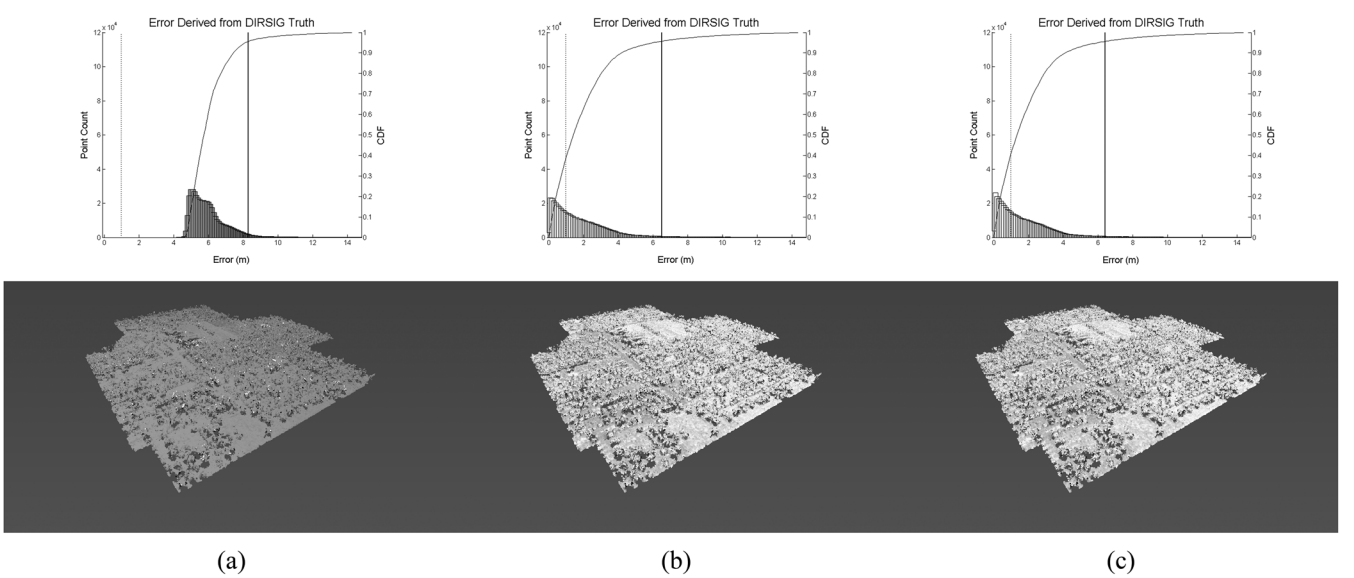
Two major sources of error are analyzed using this method. The first source of error comes from the SfM process itself. The effects of this error are isolated by using noiseless DIRSIG sensor models in the georegistration process. Error in geoaccuracy caused by the SfM process often manifests itself in poor image correspondence. Optimized image correspondence can still contain some small amount of error. This can cause small errors in camera pose estimation as well as point triangulation. The synthetic DIRSIG test imagery is taken from a nadir-looking direction, and therefore, has a very low base-to-height ratio (B/H). This low B/H causes poor image correspondence to have significantly more effect on the error in the Z-dimension.[29]

Error is calculated for each dimension X, Y, and Z, however, for illustrative purposes the error is displayed as the Euclidean distance from the truth. Given the low B/H, the error resulting from the SfM process will primarily be in the Z-dimension. However, if the georegistration transform is flawed the error will be in every dimension. Figure 7 shows the error for each method calculated using the DIRSIG noiseless sensor models.

Figure 7(a) shows a significant amount of error, caused by a poor estimation in the georegistration transform. Small errors in the relative camera pose estimation are amplified in the scale change between the relative and absolute coordinate systems. The other two methods shown in Figs. 7(b) and 7(c) show approximately the same amount of error, again almost completely in the Z-dimension. Given the random distribution of the error in the Z-dimension it is likely caused by the error from the SfM process. This result is expected when using noiseless sensor models.

5.2 *Georegistration Error Using DIRSIG Noisy Sensor Model*

A more realistic scenario would be when the sensor pose models contain a small amount of random noise. The error analysis was repeated while adding random error to



(a) (b) (c)

Fig. 7 Error is calculated using DIRSIG truth and represented as the absolute Euclidean distance from the truth. The georegistration methods are processed using the noiseless DIRSIG sensor models. Three methods are tested, (a) is the error found using the camera-center–based georegistration transform, (b) is the error found using the camera model–based transform, and (c) is simply triangulating correspondence using the sensor model. Each plot shows a histogram of error with 1 m of error marked by a dotted line. The solid line marks the 95% cumulative distribution value. The histogram's cumulative distribution function (CDF) is plotted over the histogram as well.

the sensor position and pointing information, with the amount of error determined from Applanix specifications.[28] The RMS error in position is 0.1 m in the horizontal direction, and 0.2 m in the vertical. The error in pointing is 0.015 deg for roll and pitch, and 0.040 deg for heading. Figure 8 shows the error calculated for each method.

Similar to the noiseless sensor, the camera center–based transform has a significant amount of error; the addition of noise only caused this to increase. Adding noise to the sensor information had a significant impact on directly re-triangulating each point, as seen in Fig. 8(c). A small change in pointing information from a long distance will drastically alter the projection of each pixel out of each frame, causing the triangulation solution to contain significantly greater error, as noted in Fig. 5. The augmented camera model–based transformation remained only mildly affected by the noise. This can be seen by measuring the change in position of the 95% cumulative distribution value relative to the noiseless sensor position, as shown in Table 1.

Table 1 A comparison of the 95% cumulative distribution values for each georegistration approach between noiseless and noisy sensors.

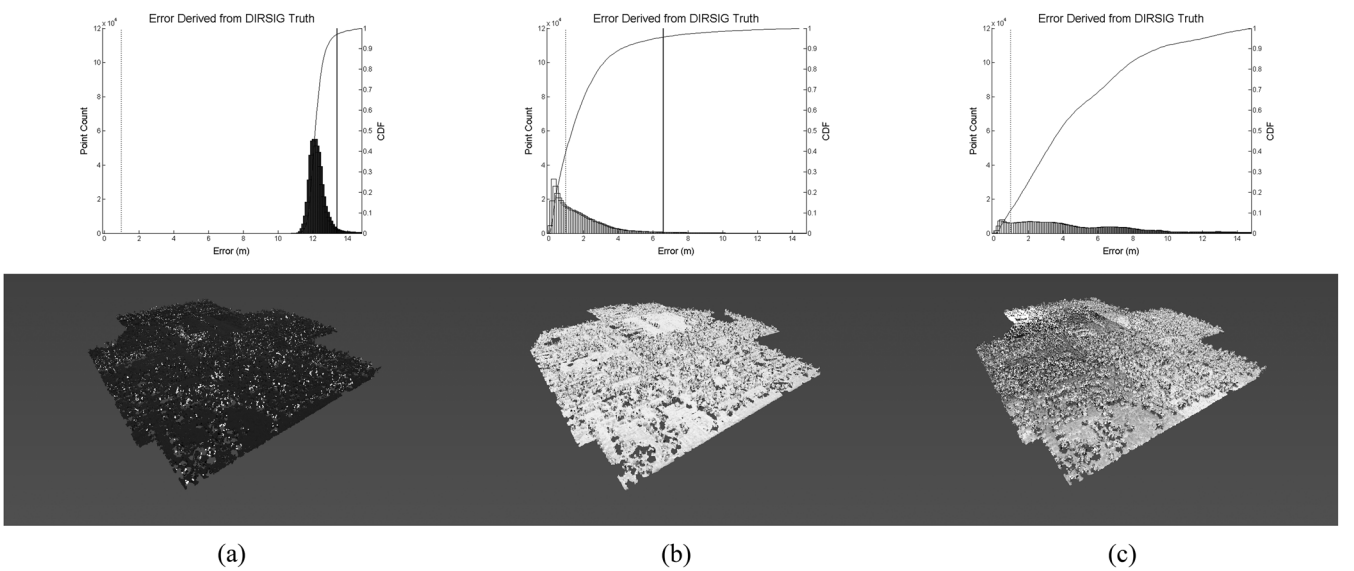
	95% cumulative distribution function (CDF) value noiseless (m)	95% CDF value noisy (m)	Change (m)
Camera centers approach	8.3	13.5	5.2
Augmented camera model approach	6.3	6.5	0.2
Direct triangulation approach	6.3	15.9	9.6

5.3 *Reducing the SfM Error*

Two major sources of error are prevalent within this error analysis. The first source of error, which originates from the georegistration process itself, is addressed in the previous sections. The second major source of error comes from the SfM process, in which image features may be inaccurately matched between images. A mismatch of a few pixels may translate into several meters of error in triangulation due to the small base-to-height ratio in near-nadir imaging. Often SfM algorithms threshold the image correspondence based on a threshold correspondence value and this threshold can be adjusted to filter the corresponding points accordingly.[16]

The algorithm used to generate these point clouds has one threshold metric which filters image correspondence based on photometric consistency using a normalized cross correlation.[30] The software uses this threshold to determine the quality of an image correspondence, with a range of −1 (bad) to 1 (good). The consistency value used in the creation of every point cloud shown in this analysis so far has been 0.8, and the error analysis for the consistency values of 0.7, 0.8, and 0.9 are shown in Fig. 9.

The histograms shown in Fig. 9 depict a trend as the consistency threshold is increased. The total number of points decreases with the increase, which can be seen by either the relative total area of each histogram or by the decreasing number of visible points in each error visualization. However, it is noted that the 95% point in each histogram, shown by the solid line, is approaching zero with the increase of the consistency threshold. This means that even though the total number of points is decreasing with the increase of threshold, the number of points with a large geographic error is decreasing. It can be concluded that the extreme



(a) (b) (c)

Fig. 8 Adding noise to the DIRSIG sensor models yields a more accurate representation of a real life scenario. The georegistration methods are processed using a noisy DIRSIG sensor model. Three methods are tested, (a) is the error found using the camera-center–based georegistration transform, (b) is the error found using the camera model–based transform, and (c) is simply triangulating correspondence using the sensor model. Each plot shows a histogram of error with 1 m of error marked by a dotted line. The solid line marks the 95% cumulative distribution value. The histogram's CDF is plotted over the histogram as well.

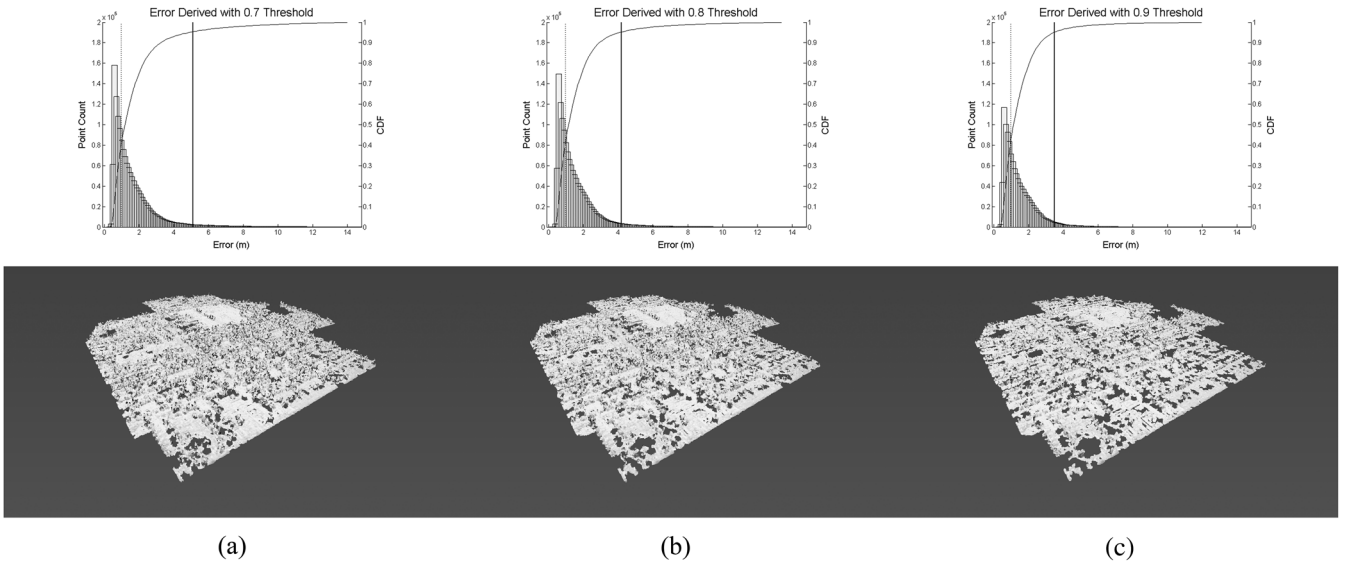


Fig. 9 Error from the SfM process often comes from mismatch in image correspondence. In the process used in this work, image correspondence is filtered using normalized cross correlation photometric consistency measurement. This figure shows error analysis done using the augmented sensor transform for different consistency threshold values. (a) Threshold 0.7 (b) Threshold 0.8 (c) Threshold 0.9.

error values in the georegistration process are likely due to triangulation error in the SfM process caused by errors in image correspondence.

6 Conclusions

Automatically extracted, accurate scene structure is the ultimate goal for many researchers who use computer vision techniques with long range aerial photogrammetry. It is, therefore, important to identify, quantify, and minimize errors that may arise during this process. This work has presented a new method for analyzing this type of data using a fully synthetic dataset. It was also shown how control over the entire imaging process can be used to isolate and analyze specific sourses of error in the SfM georegistration process. Three methods of SfM point cloud georegistration methods were analyzed to demonstrate this capability, namely the camera centers approach, the augmented camera model approach, and the direct triangulation approach. Two main sources of error were examined using this analysis approach.

The first source of error comes from the SfM process itself, often from error in image-to-image correspondence. Many factors can contribute to error in image correspondence, for example, a wide-baseline camera geometry lends itself to error as a result of the large difference in object appearance. Poor texture definition or repetitive texture on the object's surface can also lead to a mismatch. This is also the case with other imaging modalities, such as objects in thermal infrared imagery. These objects have rather poor texture and definition, which would prove challenging to most feature correspondence algorithms. An error in correspondence propagates throughout the SfM process, causing errors in camera pose estimation and scene structure triangulation. The analysis technique was also used to show that adjusting a single parameter to minimize the error in the correspondence can have an impact on the total amount of error.

The second source of error is within the georegistration transform itself, which is affected by error in the SfM process, and by error in the additional information used to generate the transform. For this work, additional information is found in the GPS/INS data. Given no error in this information, each approach was only affected by the SfM error. The SfM error in the test dataset for this work was substantial, and this led to a significant amount of the error in the camera centers approach. The augmented camera centers approach and the direct triangulation remain significantly less affected. When random instrument error is added to the GPS/INS readings, the direct triangulation approach fails. This is expected, as the instrument error directly affects the triangulation. The random error increases the total error in both the camera centers and augmented camera model approaches, however, the latter is far more robust than the former. The fact that the augmented camera model approach performed better than the camera centers approach in all tests supports the idea that geotags alone do not provide (a) the level of accuracy obtained by including an additional error minimization through trianguation or (b) enough data to overcome this lack of error minimization.

One general observation can be stated from the analysis presented in this work. The error caused by GPS/INS was shown to have a very significant impact on some georegistration methods, but could also be greatly mitigated using other methods. The error coming from the SfM process was persistent and difficult to reduce across all the methods tested. Feature matching algorithms other than SIFT may produce better results, depending on the image content. Along the same line, other algorithms which derive image-based point clouds may introduce less error into the georegistration process. For improving the georegistration accuracy of point clouds, focus should be put on minimizing the error from the point cloud extraction process, as error introduced by the GPS/INS system can be easily lessened.

The fully synthetic dataset permits error analysis of point cloud georegistration algorithms, although our dataset is not

without fault; the synthetic texture can be homogeneous in both color and texture. This is a challenge to many image correspondence algorithms, and consequently the SfM process. However, having an absolute position truth for each pixel provides an analysis tool that would otherwise be unavailable, and other georegistration methods may benefit from the analysis this sort of dataset could provide. A similar, but larger, dataset to the one used here can be found at Ref. 31.[21] The accuracy of other georegistered SfM methods could be analyzed using the analysis method presented in this work, providing a greater understanding of the performance of these existing algorithms.

Acknowledgments

Portions of this work were carried out using funding received from the United States Department of Energy under BAA PDP08 Grant Number DE-AR52-07NA28115.

References

1. R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, Cambridge, United Kingdom (2008).
2. J. Philip, "A non-iterative algorithm for determining all essential matrices corresponding to five point pairs," *Photogramm. Record* **15**(88), 589–599 (1996).
3. N. Snavely, "Bundler: structure from motion (sfm) for unordered image collections," http://phototour.cs.washington.edu/bundler/ (12 October 2013).
4. F. Neitzel and J. Klonowski, "Mobile 3D mapping with a low-cost UAV system," in *Proc. Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXVIII-1/C22*, pp. 39–44 (2011).
5. D. Crandall et al., "SfM with MRFs: discrete-continuous optimization for large-scale structure from motion," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(12), 2841–2853 (2013).
6. B. P. Hudzietz and S. Saripalli, "An experimental evaluation of 3D terrain mapping with an autonomous helicopter," in *Proc. Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXVIII-1/C22*, 137–142 (2011).
7. A. Koutsoudis et al., "Multi-image 3D reconstruction data evaluation," *J. Cultural Heritage* **15**(1), 73–79 (2014).
8. N. Snavely, *Scene Reconstruction and Visualization from Internet Photo Collections*, PhD Thesis, University of Washington (2008).
9. M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. of the ACM* **24**(6), 381–395 (1981).
10. C. Wang, K. Wilson, and N. Snavely, "Accurate georegistration of point clouds using geographic data," in *Proc. Int. Conf. on 3DTV*, pp. 33–40, IEEE (2013).
11. A. Wendel, A. Irschara, and H. Bischof, "Automatic alignment of 3D reconstructions using a digital surface model," in *Proc. Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 29–36, IEEE (2011).
12. D. Walvoord et al., "Geoaccurate three-dimensional reconstruction via image-based geometry," *Proc. SPIE* **8747**, 874706 (2013).
13. D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Computer Vision* **60**(2), 91–110 (2004).
14. B. Triggs et al., "Bundle adjustmenta modern synthesis," in *Vision Algorithms: Theory and Practice*, pp. 298–372, Springer, Berlin (2000).
15. Y. Furukawa, "Clustering Views for Multi-view Stereo (CMVS)," http://grail.cs.washington.edu/software/cmvs/ (10 November 2013).
16. Y. Furukawa and J. Ponce, "Patch-based Multi-view Stereo software (PMVS - Version 2)," http://grail.cs.washington.edu/software/pmvs/ (10November2013).
17. D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(6), 756–770 (2004).
18. Y. Furukawa et al., "Towards internet-scale multi-view stereo," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1434–1441, IEEE (2010).
19. S. Brown, "Dirsig: The Digital Imaging and Remote Sensing Image Generation model," http://dirsig.org/ (8 January 2014).
20. E. Ientilucci and S. Brown, "Advances in wide-area hyperspectral image simulation," *Proc. SPIE* **5075**, 110–121 (2003).
21. K. Salvaggio and C. Salvaggio, "Automated identification of voids in three-dimensional point clouds," *Proc. SPIE* **8866**, 88660H (2013).
22. J. McGlone et al., Manual of Photogrammetry, Bethesda, Maryland, American Society for Photogrammetry and Remote Sensing (2004).
23. K. Arun, T. Huang, and S. Blostein, "Least-squares fitting of two 3-D point sets," *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-9**(5), 698–700 (1987).
24. R. Jain, R. Kasturi, and B. Schunck, *Machine Vision*, Vol. 5, McGraw-Hill, New York, NY (1995).
25. W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallogr. Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **32**(5), 922–923 (1976).
26. W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* **34**, 827–828 (1978).
27. N. Snavely, S. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections in 3D," *ACM Trans. on Graphics (TOG)* **25**(3), 835–846 (2006).
28. Applanix., "POSTrack specifications," applanix.com (19 January 2014).
29. P. Wolf and B. Dewitt, *Elements of Photgrammetry with Applications in GIS*, McGraw-Hill, New York, NY (2000).
30. Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(8), 1362–1376 (2010).
31. K. Salvaggio, "Synthetic imagery and truth data for evaluating multiview 3D extraction algorithms," http://dirsapps.cis.rit.edu/3d-dirsig-truth/ (19 January 2014).

David Nilosek is an imaging science PhD student at the Rochester Institute of Technology. He received his imaging science BS degree from RIT in 2008. He is a student in the Digital Imaging and Remote Sensing Laboratory in the Chester F. Carlson Center for Imaging Science. His current research interests include three-dimensional model extraction from aerial imagery, and computer vision-based applications to three-dimensional point clouds. He is a student member of SPIE.

Derek J. Walvoord received his BS and PhD degrees in imaging science from Rochester Institute of Technology's Chester F. Carlson Center for Imaging Science in 2002 and 2008, respectively. In the summer of 2008, he joined Exelis (formerly ITT Space Systems Division) as a senior image scientist. He teaches externally as an adjunct professor at RIT. His professional interests include spatial image processing, linear systems and computational imaging, and photogrammetric sensor modeling.

Carl Salvaggio is a professor of the Imaging Science in the Chester F. Carlson Center for Imaging Science at RIT. He is a member of the Digital Imaging and Remote Sensing Laboratory, teaching and conducting research in digital image processing, remote sensing, and computer science. His research interests lie in thermal infrared phenomenology, exploitation, and simulation; three-dimensional geometry extraction from multiview imagery and LiDAR data; and still and motion image processing for various applications.