Automatic Registration of LiDAR and Optical Imagery using Depth Map Stereo

Hyojin Kim
Lawrence Livermore
National Laboratory
kim63@llnl.gov

Carlos D. Correa*
Google Inc.
cdcorrea@google.com

Nelson Max
University of California
Davis
max@cs.ucdavis.edu

Abstract

Automatic fusion of aerial optical imagery and untextured LiDAR data has been of significant interest for generating photo-realistic 3D urban models in recent years. However, unsupervised, robust registration still remains a challenge. This paper presents a new registration method that does not require priori knowledge such as GPS/INS information. The proposed algorithm is based on feature correspondence between a LiDAR depth map and a depth map from an optical image. Each optical depth map is generated from edge-preserving dense correspondence between the image and another optical image, followed by ground plane estimation and alignment for depth consistency. Our two-pass RANSAC with Maximum Likelihood estimation incorporates 2D-2D and 2D-3D correspondences to yield robust camera pose estimation. Experiments with a LiDAR-optical imagery dataset show promising results, without using initial pose information.

1. Introduction

Photo-realistic 3D scene representations are widely used in aerial motion imagery systems, GIS, and other 3D mapping applications. For a large-scale 3D model of building and terrain environments, LiDAR-based model acquisition has recently received attention due to its robustness and cost effectiveness, in comparison with manual 3D modeling. LiDAR approaches outperform multi-view stereo reconstructions in terms of accuracy, though multi-view stereo is also cost effective for 3D model acquisition. One drawback in this approach is that most LiDAR data does not provide color or texture information; it requires further processing such as LiDAR-optical imagery registration for a photo-realistic textured model. The LiDAR-imagery registration consists of determining the camera parameters (*i.e.*, projection matrix) for each optical image, so that the image is projected onto the target LiDAR data.

*This work was done while Dr. Correa worked at Lawrence Livermore National Laboratory. He is currently affiliated with Google Inc.

A number of registration algorithms have been proposed in recent years; however, most approaches require initial camera parameters from GPS and Inertial Navigation System (INS) information. Methods using 2D/3D line features have difficulty in registering images that do not contain obvious line segments. 3D-3D registration approaches are costly to compute though they do not require initial camera parameters. An inexpensive, unsupervised, robust, automatic registration algorithm that does not require prior knowledge of structure or camera parameters is needed.

In this paper, we present a new approach to register 3D LiDAR data with 2D optical imagery. The proposed method does not require initial camera parameters, but requires two optical images used to generate a stereo depth map. Instead of directly reconstructing a 3D model from each stereo depth map (which is typically inferior to LiDAR data), we run feature correspondence between LiDAR and stereo depth maps, in order to match a few of the most distinctive features for correct registration. Our depth map-based registration is an improvement over approaches that rely mostly on certain types of geometric features (typically line segments), and over approaches that use computationally expensive 3D-3D registration.

To guarantee depth (brightness) consistency between LiDAR and stereo depth maps, ground plane estimation and alignment are performed for oblique imagery. Ground plane-based depth maps in which the brightness is based on the height give more reliable feature correspondence. Our proposed two-pass RANSAC scheme with Maximum Likelihood (ML) estimation accurately estimates camera parameters. We tested the proposed method with a LiDAR-optical imagery dataset [6] covering Rochester, NY.

2. Related Work and Contribution

Previously published papers can be classified into two categories; 2D-3D registration using object features; and 3D-3D registration using Structure from Motion (SfM).

A number of algorithms are based on 2D-3D registration using feature detection from both 3D LiDAR and 2D optical imagery. The features are collected by detecting specific

points or regions of interest (*e.g*., building corners, facades) and camera parameters are estimated so that the features from LiDAR are consistent with those from 2D images, given the camera pose. These algorithms start with initial camera parameters, obtained most often from a GPS/INS system, and then refine them.

Several papers focused on registration of ground-level LiDAR and optical imagery [8, 9, 10, 15, 16, 23, 24]. The algorithm of Lee *et al*. [8] detects vanishing points from straight lines in both LiDAR and images, and then does pose estimation by decoupling camera rotation and translation. Stamos and Allen [15] used building facades to match rectangles. To compute an optimal camera transformation, Liu and Stamos [9, 10] used features from vanishing point extraction (at least two vanishing points) and a matching of the 2D features with 3D features in LiDAR data that maximizes an overlap measure. They also proposed a user-interface for texture mapping of 2D images onto 3D LiDAR data at interactive rates [10]. Wang *et al*. [23] used a similar approach, relying on vanishing points. These methods are not suitable for scenes where parallel lines to induce vanishing points are not easily detectable. Yang *et al*. [24] used feature matching to register images from a hand-held camera with range data or color imagery. Stomas *et al*. [16] improved their previous algorithms by incorporating line segment matching.

There are several papers that register different types of ground-level imagery. The algorithm of Troccoli and Allen [19] uses shadow matching to register 2D images to a 3D model. Kurazume *et al*. [7] used edge matching, and an M-estimator to register laser-scanned ground-based objects (*e.g*., a Buddha statue).

Similar techniques using feature matching to handle aerial imagery have also been proposed. Frueh *et al*. [4] used line segment matching to adjust initial camera parameters from GPS/INS by exhaustively searching camera position, orientation, and focal length. The method of Vasile *et al*. [21] generates pseudo-intensity images with shadows from LiDAR to match 2D imagery. Then camera parameters from GPS and camera line of sight information are exhaustively estimated, similar to [4]. The algorithm of Ding *et al*. [2] uses vanishing points for oblique aerial imagery to extract features called 2D Orthogonal Corners (2DOCs). Initial camera parameters are then refined using M-estimator RANSAC. The algorithm of Wang and Neumann [22] does line segment detection and matching, followed by a two-level RANSAC to divide putative feature matches into multiple groups. They introduced 3 Connected Segments (3CSs) that they claimed are more distinctive than 2DOCs. Mastin *et al*. [13] gave a statistical approach using mutual information between LiDAR and oblique optical imagery. The algorithm uses 3D-2D rendering of height and probability of detection (pdet) attributes.

These 2D-3D approaches using feature detection have several limitations. First, they do not work when LiDAR data provide no camera information, due to their heavy dependence upon initial camera parameters from GPS, INS, compass, and other measurements. Also, they rely mainly on line segments that are detected from buildings and other man-made structures. Thus they are not suitable for images that contain few specific line segments.

Another approach is based on 3D-3D registration using SfM. Zhao *et al*. [25] proposed an algorithm that reconstructs 3D geometry from video sequences and registers it with LiDAR data using the Iterative Closest Point (ICP) algorithm. Liu *et al*. [11] used SfM to generate a 3D point cloud and registered it with LiDAR data. Methods in this category do not typically require a priori knowledge such as GPS/INS information. Compared to 2D-3D registration, however, 3D-3D registration is more difficult and it also requires more accurate 3D multi-view reconstruction.

The novelty of our approach lies in the characteristics of our algorithm, that takes advantage of both approaches. Like the 3D-3D registration approach, our method does not require initial camera pose information. Thus our method can be used for a wider set of scenes without GPS (*e.g*., indoor images). A 2.5D depth map is a 2D projected raster image with a color at each pixel indicating height above a ground plane. Our feature correspondence between 2.5D depth maps gives flexibility as well as less computational cost, compared to 3D feature matching. With underlying 3D information in 2.5D LiDAR depth maps, the proposed matching scheme uses two kinds of matching criterion by incorporating 2D-2D and 2D-3D correspondences, which increases the robustness of our approach. More importantly, our method overcomes the limitations on the feature selection of the existing 2D-3D approaches by detecting any feature available in the imagery (*e.g*., vegetation).

In addition, we propose several techniques to improve registration accuracy. The edge-preserving dense correspondence used to generate stereo depth maps enables us to detect as many distinctive features (at or near object boundaries) from the depth maps as possible, whereas many existing algorithms tend to excessively smooth out edges and corners to suppress matching errors (mostly due to occlusion and textureless regions). We also propose a robust two-pass RANSAC scheme with a likelihood estimator that utilizes two types of matching criterion (epipolar and projection constraints). The two-pass modified RANSAC allows more accurate camera estimation, even in the case of a small number of inliers with numerous outliers.

3. Algorithm

Now let us describe our novel automatic registration algorithm for aerial LiDAR and optical imagery. For 3D LiDAR imagery, any LiDAR dataset where each point has an

x, y, and z value can be used. To generate stereo depth maps from optical imagery, we assume intrinsic camera parameters (*e.g.*, focal length) are known. In the case of unknown parameters, we estimate the information via camera calibration and pose estimation between the optical images.

The registration process starts by generating depth maps from both LiDAR and optical imagery (Subsection 3.2). Depth map stereo (feature correspondence) between two depth maps gives a set of matched feature pairs, simplifying the problem into 2D-3D correspondence (Subsection 3.3). Then, a subsequent two-pass RANSAC is performed to estimate a camera matrix, along with removing a number of outliers from the matched pairs (Subsection 3.4). We use a modified RANSAC, known as MLESAC [18], which maximizes the likelihood of the correspondences between two depth maps. Finally we use texture mapping from the optical image onto a triangulation of the LiDAR data to generate a photo-realistic 3D model.

3.1. Model

The goal of the registration algorithm is to find a camera projection matrix for each view of optical imagery so that each optical image is properly mapped onto the LiDAR model. In the depth map stereo stage, image features are matched between a LiDAR depth map and a stereo depth map from optical imagery. Since the LiDAR depth map incorporates underlying 3D coordinates, the main problem simply becomes 2D-3D correspondence. Let $\mathbf{P}$, $\mathbf{x}$, $\mathbf{X}$ denote, respectively, a camera projection matrix for an optical image to be registered, a set of 2D homogeneous points in the image, and a set of 3D homogeneous points. Our goal is to find $\mathbf{P}$ such that,

$$\mathbf{x} = \mathbf{P}\mathbf{X} \tag{1}$$

where $\mathbf{P}$ is 3×4 matrix that involves intrinsic camera parameters (3×3 $\mathbf{K}$ matrix) and extrinsic camera parameters (3×4 $[\mathbf{R}|\mathbf{t}]$ matrix). The intrinsic camera parameters include focal length, principal point, skew, etc., specifying the camera optics and sensor. The extrinsic camera parameters are decomposed into 3×3 rotation matrix $\mathbf{R}$ (camera orientation) and the translation column vector $\mathbf{t}$ (camera position). Then we want to find an optimal $\mathbf{P}$ such that the sum of reprojection error is minimized as,

$$\arg\min_{\mathbf{P}} \sum_{i=1\ldots n} d(\mathbf{x}_i, \mathbf{P}\mathbf{X}_i) \tag{2}$$

where n is the number of matched feature pairs, and d is the reprojection error, the Euclidean distance (in pixels) between a 2D matched feature point and a reprojected point from the corresponding 3D point. $\mathbf{P}$ can be found by using Direct Linear Transformation (DLT) with normalization of the 2D image points, or iterative non-linear optimization techniques such as Levenberg-Marquardt [5].

One challenge is that input data in the 2D-3D correspondence (the matched feature pairs) is seriously corrupted by numerous outliers, due to high ambiguity in the 2.5D features. Thus, incorrectly matched pairs from the depth map stereo stage should be properly filtered out. We present a two-pass RANSAC approach that removes outliers using a different criterion in each pass, which we discuss in Subsection 3.4. The original RANSAC [3] relies heavily on choosing a distance threshold to determine inliers. In the case of unknown distribution of outliers or too high a threshold, the estimation works poorly. We therefore adopt a modified RANSAC estimator, MLESAC [18], to find a solution that maximizes the likelihood of the correspondences. In our case, the distance error (reprojection error) is represented as a mixture model of the Gaussian and uniform distributions, with respect to inlier and outlier costs. Assuming $\sigma = 1$ (in a multivariate normal distribution), the likelihood that a given matched pair is an inlier Pr_{in} is:

$$Pr_{in} = m\Big(\frac{1}{\sqrt{2\pi}}\Big)^2 \exp\Big(\frac{-d^2}{2}\Big) \tag{3}$$

where m is the mixing number (in $[0, 1]$). Also the likelihood that a given pair is an outlier Pr_{out} is:

$$Pr_{out} = (1 - m)\frac{1}{\nu} \tag{4}$$

where ν is a pre-computed constant (maximum distance of all point pairs in image domain). The mixing number m is initially 0.5 according to [18] and is updated in an iterative manner using Expectation Maximization (EM) such that:

$$m = \frac{1}{n}\sum_{i\ldots n} \frac{Pr_{in}}{Pr_{in} + Pr_{out}} \tag{5}$$

The goal is to choose an optimal $\mathbf{P}$ that minimizes the cost (maximizing the likelihood), which is represented as the negative log likelihood:

$$-L = -\sum_{i=1\ldots n} \log\left(Pr_{in} + Pr_{out}\right) \tag{6}$$

Together with the projection constraint in Equation 2, we also use the epipolar constraint in 2D-2D image correspondence between two depth maps. Given an estimated Fundamental matrix, its epipolar constraint determines the validity of each matched pair, that is, a matched point should be searched for along the epipolar line on which the reference point is located. We use the epipolar constraint in the first pass of our RANSAC scheme, followed by the projection constraint in the second pass. See [5] for more details on the Fundamental matrix and the epipolar constraint.

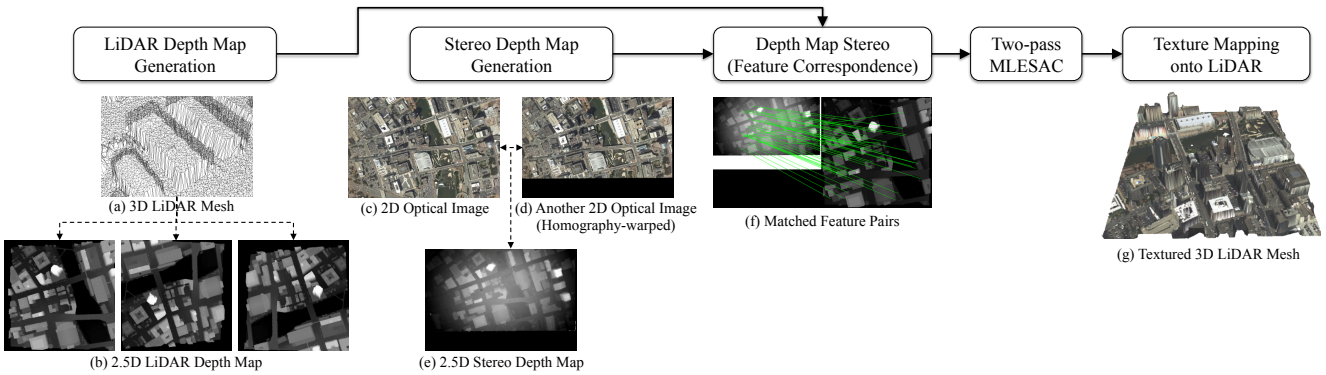

Figure 1. An overview of our algorithm.

3.2. Depth Map Generation

3.2.1 LiDAR Depth Map

In the first stage, we generate 2.5D depth maps (height maps) from both LiDAR and optical imagery. We first construct a LiDAR mesh from the input LiDAR data using 2D Delaunay triangulation of the x and y coordinates (Figure 1 (a)). Then a gray-scale color is assigned to each vertex in the mesh, depending on the height value (z value in the Rochester dataset [6]), which converts the mesh into a height map. The color (brightness) assignment is relative to the maximum and minimum height/elevation values, to obtain a ground-based depth map. Then we create multiple depth maps, each of which is rendered from a different camera position and direction, having different occlusions (Figure 1 (b)). The reason for preparing multiple depth maps is due to uncertainty of occlusions in each input optical image. From our experiments, $4-8$ LiDAR depth maps are sufficient. Every pixel in each LiDAR depth map stores an interpolated 3D point, which is later used for matching validity (Subsection 3.3) and 2D-3D correspondence (Subsection 3.4).

3.2.2 Stereo Depth Map

To create a depth map for each optical image, we perform dense correspondence between the source image to be registered (Figure 1 (c)) and another target image which is homography-warped to the source image (Figure 1 (d)), with a baseline large enough to provide accurate correspondences. In aerial imagery, we found that images separated by between 5 and 22.5 degrees work best. The key point in the dense correspondence is to preserve edges and other significant features that give uniqueness of regions. To achieve these goals, we have developed an edge-preserving dense correspondence based on multi-resolution disparity propagation and bilateral filters.

A dense disparity map is found by iterating over three local operators: *search*, *propagation* and *affine smoothing*. We define disparity as the vector difference between a pixel in the source image and a corresponding pixel in the target image. We use normalized cross correlation to measure the fit of a disparity. *search* finds the best match between a 5×5 pixel patch of the source image and neighboring patches of the target image. *propagation* refines the disparity at a given pixel by propagating the disparities of neighboring pixels in the source image. If using one of the neighbors' disparity results in a better fit, the algorithm updates the disparity of the pixel with that of the neighbor. The repeated execution of these two steps effectively performs a walk around the image to find a good match, without the need to increase the size of the search or propagation patches. *affine smoothing* fits an affine transformation to the disparities in a patch around each pixel to keep the disparities structured. These three operators are iterated for a number of times $(10-20)$ or until the propagation converges to a solution where there is no change in disparity, and is repeated at multiple scales on a Gaussian pyramid of the image pair. In order to guarantee depth discontinuities at or near object boundaries (buildings), we introduce a bilateral filter [17] for both the affine smoothing and propagation.

A depth map can be obtained from the disparity via epipolar geometry estimation (Fundamental matrix) followed by camera pose estimation (*e.g.* RANSAC with 8-point algorithm [5]). In the case of oblique imagery, however, the depth map may have inconsistent brightness with the ground-based LiDAR depth map. Depth (brightness) consistency between two depth maps is critical in the following feature correspondence. In order to generate a ground-based stereo depth map, we perform ground plane estimation and alignment by finding a dominant plane of the reconstructed scene using RANSAC-based plane fitting. Then all the reconstructed 3D points together with the estimated camera pose are transformed so that the ground plane is axis-aligned (xz-plane) where y-axis represents the height. This axis-aligned reconstruction gives a ground-

based depth map, as shown in Figure 1 (e).

3.3. Depth Map Stereo

Depth map stereo, not to be confused with the previous stereo process to produce a depth map, consists on finding a small set of the most distinctive feature correspondences between two depth maps, in our case a LiDAR depth map and the stereo depth map from an optical image pair (Figure 1 (f)). Given a depth map pair, we use sparse feature matching, invariant to scale and orientation, such as Scale Invariant Feature Transform (SIFT) [12] and Speeded Up Robust Feature (SURF) [1]. We believe that an invariant feature from 2.5D depth maps can be interpreted as a geometric uniqueness in 3D. Our experiments show that 2.5D depth map feature matching is robust if the brightness of the depth maps are fairly consistent.

Experimenting with several algorithms, we found that SURF offered fast, accurate matching and computational flexibility. SURF also showed robustness against the relatively assigned colors in LiDAR depth maps. Due to blurred and obscured depth map images (compared to typical optical imagery), we adjust SURF parameters such as the number of pyramid octaves, the number of layers within each octave, and the Hessian threshold to maximize the matching accuracy. In particular, we use lower Hessian thresholds $(100 - 200)$ to detect as many features as possible.

As discussed previously, multiple LiDAR depth maps of the same region are generated. If there are n LiDAR depth maps, we perform n depth map stereo processes for each stereo depth map. The multiple LiDAR depth maps offer several advantages. First, they overcome a potential occlusion problem; some features from the stereo depth map are possibly occluded in some of the LiDAR depth maps, which may lead to incorrect matching. Second, they provides matching validity, that is, if a feature point in the stereo depth map has inconsistent matched points in the LiDAR depth maps by checking the associated 3D points, it is discarded. Third, it removes the need for manual initial registration of the LiDAR and imagery and thus makes it more appropriate for fully automated registration.

Two scenarios can be applied to this stage. One is to apply $1 : n$ matching and extract valid matched pairs by checking matching inconsistency, described above. These matched pairs are used for 2D-3D correspondence. Another scenario is to choose a LiDAR depth map that gives more valid matched pairs than any other LiDAR depth map and to perform depth map stereo between the stereo depth map and the chosen LiDAR depth map. According to our experiments, both scenarios have provided similar results.

3.4. Two-pass MLESAC

Because every LiDAR point has underlying 3D information, the problem simply becomes 2D-3D correspondence.

Figure 2. Two registration result pairs between two-pass RANSAC (*top*) and two-pass MLESAC (*bottom*). The colors from the optical images were modified by lighting and shading to reveal the LiDAR shapes.

One key issue is matching ambiguity, resulting in too many outliers, probably due to insufficient texture and gradient information in both depth maps. We therefore perform two-pass MLESAC to acquire an optimal camera projection matrix from such noisy matching data.

The previous depth map stereo incorporates 2D-2D correspondence (2D feature matching between two depth maps) and 2D-3D correspondence (using underlying 3D points in the LiDAR depth map), which provides two types of matching criteria (epipolar and projection constraints). In the first pass, we use the epipolar constraint to identify outliers in 2D-2D correspondence. A matched pair that does not lie in the same epipolar line (within $2 - 3$ pixels) becomes an outlier, given an estimated fundamental matrix. The second pass uses the projection constraint in 2D-3D correspondence, that is, any matched pair that has a reprojection error (more than $1 - 2$ pixels) is discarded.

As discussed earlier, MLESAC is more robust than RANSAC against numerous outliers and an undetermined distance threshold. Figure 2 shows two results improved by MLESAC, compared to the conventional RANSAC. The numbers of RANSAC and EM iterations are 1000 and 3, respectively.

Once a projection matrix for each stereo depth map is estimated, we map the original image onto the LiDAR mesh (Figure 1 (g)). In the case of multiple images mapping to the same region, the colors can be a weightedly average, with weigths depending on the normal and the viewing direction.

4. Experimental Results

To evaluate the effectiveness of the proposed algorithm, we performed experiments with aerial LiDAR and optical imagery from [6], covering a portion of downtown

Figure 3. Registration results. From *left* to *right*, the untextured 3D LiDAR model (height map), a manually registered 3D model, and a 3D model generated by our registration process. Lighting and shading added, as in Figure 2.

Table 1. Quantitative evaluations of the LiDAR-imagery registration. In the first evaluation, n_d, d_m, and d_r indicate the number of manually labeled correspondences, the average reprojection error using the manually estimated camera pose, and the average reprojection error of our estimated camera pose, respectively. In the second evaluation, n_s and s_r are the number of total vertices and the similarity (%), respectively.

	n_d	d_m	d_r	n_s	s_r
Site 01	12	0.99	3.84	286,954	89%
Site 02	13	0.84	1.75	340,277	93%
Site 03	12	1.42	4.21	349,779	88%
Site 04	13	1.53	2.95	340,640	90%
Site 05	10	1.58	1.77	193,241	89%
Site 06	12	2.13	1.27	152,664	93%
Site 07	18	0.83	5.06	288,416	91%
Site 08	13	0.69	1.07	267,886	92%
Site 09	12	0.93	2.67	239,878	91%
Site 10	12	1.10	2.68	187,154	93%

Rochester in NY (43.155° N, 77.606° W). The LiDAR imagery covers approximately 2.5 km × 1 km, consisting of 40 tiled LiDAR data files. Since each optical image (1000 × 688) covers about 2 × 2 tiles, we combined 4 LiDAR data tiles into a single polygonal mesh so that one image is registered with each mesh. In contrast to the LiDAR data, optical imagery covers a much larger area. We therefore selected optical images (about 40 out of 681 images) covering the area where LiDAR data was available.

For each site (2 × 2 tiles), we performed our registration process in order to register a LiDAR mesh with an optical image. In the depth map generation stage, we used Bundler [14] to estimate a camera pose between the source optical image to be registered and another target image.

Two quantitative evaluations were performed to measure the quality and the correctness of the finally registered 3D LiDAR model. Since no ground-truth information is provided, we instead generated a manually registered mesh model. First we labeled $10 - 20$ points in each LiDAR depth map and then found their correspondences from the optical image to be registered. Given the manually labeled

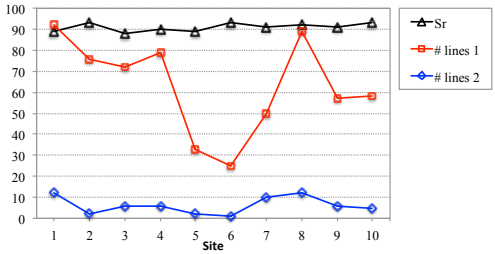

Figure 4. The similarity (s_r in Table 1) and the number of line segments in each site (Site 01 - Site 10). Lines 1 and 2 indicate salient line segments whose length is larger than 50 and 100 (in pixels), respectively.

correspondences, a camera pose for the optical image is computed. Figure 3 shows the original 3D LiDAR mesh model without textures (a height map), the manually registered model, and the model by our automatic registration process. Since the optical imagery is taken from almost directly above, the registered model does not provide details in the building walls and other vertical textures.

The first evaluation is to compute the distance error of our registration by reprojecting 3D points of the manually labeled correspondences onto the optical image and by measuring the pixel distance between the reprojected and the manually labeled points, as shown in Table 1.

The next evaluation is to compute similarity between the manually and the automatically registered mesh models. For each vertex in the mesh model, neighboring colors in the optical image onto which the vertex is projected are bilinearly interpolated. We then compute the mean of the absolute color differences of all vertices between the two mesh models (c_r). The color similarity (s_r in Table 1) is $1 - c_r$, expressed as a percent.

We also measured the running time for each registration process. Each optical depth map generation using our dense correspondence takes 130 secs, which needs to be improved. The feature detection and matching (depth map stereo) take 193 ms and 494 ms, respectively. The two-pass MLESAC with 5000 iterations takes 276 ms.

The experiments show that our automatic registration

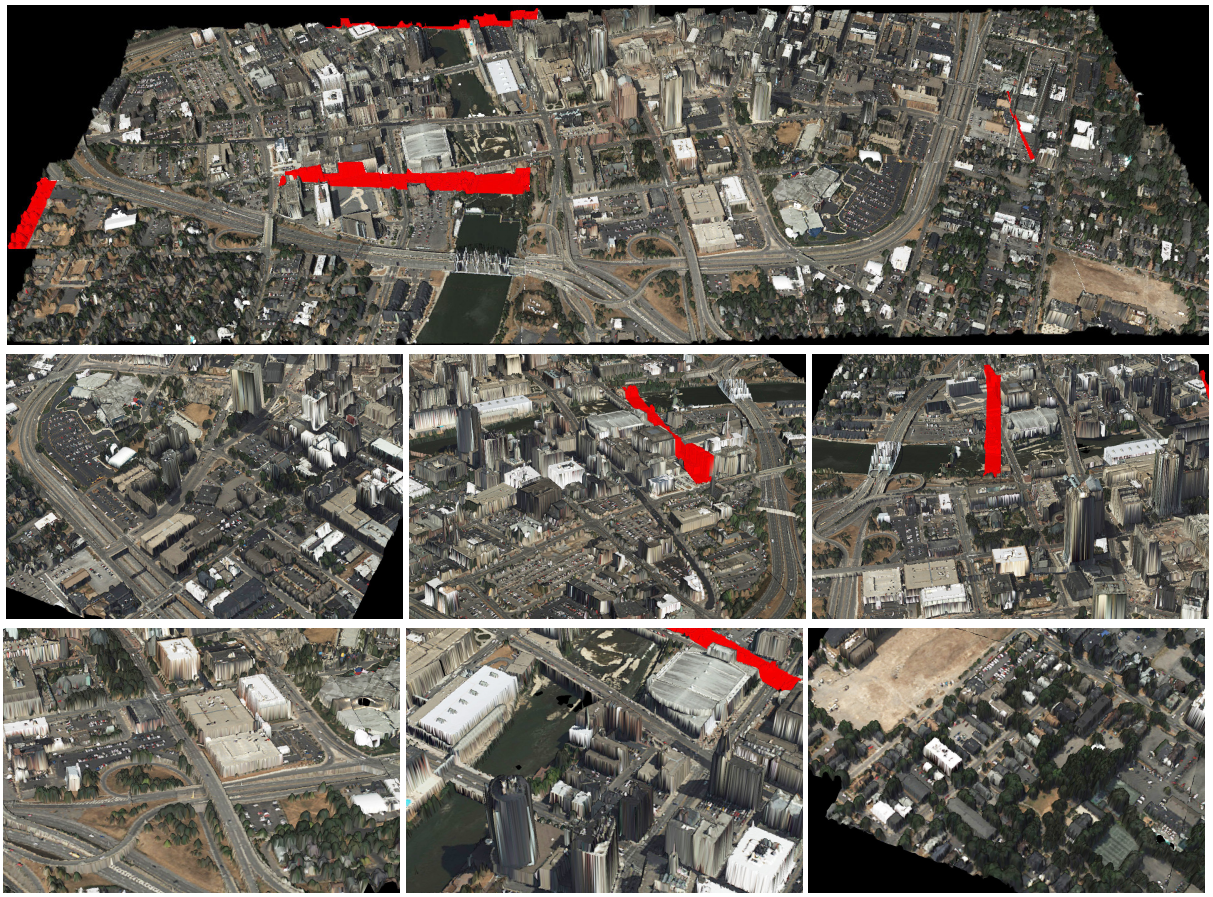

Figure 5. A registered 3D LiDAR mesh model of Rochester, NY [6]. Red color indicates no image information available in that area.

gives promising results in most regions, in comparison with the manually registered ones. The reprojection errors using the automatically registered camera pose are slightly larger than those using the manually estimated camera pose. Yet the two mesh models are visually identical because the mean color similarity is over 90%. Figure 5 shows several screen shots of the final automatically registered 3D textured model. Also, our registration approach works well even if no initial camera information is provided. Unlike most state-of-art registration algorithms that refine the given initial camera pose from GPS/INS, our method recovers a reasonably accurate camera pose without using any prior knowledge.

Furthermore, our method is suitable for regions in which there is no man-made structure. Sufficient salient lines in optical imagery are essential for most 2D-3D registration algorithms discussed in Section 2. Regardless of line segments in optical imagery, our method is still effective, due to the characteristics of the feature detection and matching. Figure 4 shows another evaluation to measure the color similarity (s_r in Table 1; roughly representing registration accuracy) and the number of salient line segments in each site. We used a line segment detector [20] to detect salient line segments. As shown in Figure 4, several sites containing a lot of trees and bushes with few salient lines (*e.g.*, site 05 and 06) are successfully registered.

The experiments also address potential errors in the manual registration. The first evaluation indicates that reprojection errors using the manually registered model and its camera pose are not negligible (d_m in Table 1). Also, the time-consuming manual registration is not as efficient as automatic registration. Nevertheless, a manual registration is a useful tool to measure the accuracy of the algorithm in the case of no ground-truth information, as has been shown in previous literature [22]. One may visually examine the quality of the registered model. However, choosing an appropriate scoring metric is subjective and difficult. Also, we observed a kind of circular gradient in the disparity and depth maps from the optical imagery (see the brighter center in Figure 1 (e)), which we think is related to a radial distortion. One may need to undistort the optical images or divide each image into multiple sub-images for which an optimal camera pose is estimated. The errors in the manual registration may also be related to this issue.

5. Conclusion

We have presented a novel registration method for aerial LiDAR and optical imagery. Our approach is based on feature matching between 2.5D depth maps, utilizing 2D-2D and 2D-3D correspondence. The proposed two-pass MLESAC where each pass employs a different matching constraint provides an optimal camera pose, effectively removing a large number of outliers. Unlike existing 2D-3D registration approaches, the proposed algorithm does not require initial camera parameters, but gives accurate, efficient registration results. This algorithm is also suitable for other registration problems where no initial pose is provided and/or no straight lines exist (*e.g.*, a laser-scanned 3D face model with 2D imagery).

Acknowledgements

This work was performed in part under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

[1] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. SURF: Speeded up robust features. In *Computer Vision and Image Understanding*, volume 110, pages 346–359, 2008.

[2] M. Ding, K. Lyngbaek, and A. Zakhor. Automatic registration of aerial imagery with untextured 3d lidar models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[3] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24:381–395, 1981.

[4] C. Frueh, R. Sammon, and A. Zakhor. Automated texture mapping of 3d city models with oblique aerial imagery. In *3DPVT*, pages 396–403, 2004.

[5] R. Hartley and A. Zisserman. *Multiple View Geometry in computer vision*. Cambridge University Press, Cambridge, UK, second edition edition, 2003.

[6] R. Institute of Technology. 3D - Rochester. Website, 2012. available at `http://dirsapps.cis.rit.edu/3d-rochester/index.html`.

[7] R. Kurazume, K. Nishino, M. D. Wheeler, and K. Ikeuchi. Mapping textures on 3d geometric model using reflectance image. In *Syst. Comput*, volume 36, pages 92–101, 2005.

[8] S. C. Lee, S. K. Jung, and R. Nevatia. Automatic integration of facade textures into 3d building models with a projective geometry based line clustering. *Comput. Graph. Forum*, 21, 2002.

[9] L. Liu and I. Stamos. Automatic 3d to 2d registration for the photorealistic rendering of urban scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 137–143, 2005.

[10] L. Liu and I. Stamos. A systematic approach for 2d-image to 3d-range registration in urban environments. In *International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

[11] L. Liu, I. Stamos, G. Yu, G. Wolberg, and S. Zokai. Multiview geometry for texture mapping 2d images onto 3d range data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2293–2300, 2006.

[12] D. G. Lowe. Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1150–1157, 1999.

[13] A. Mastin, J. Kepner, and J. F. III. Automatic registration of lidar and optical images of urban scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2639–2646, 2009.

[14] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *International Journal of Computer Vision*, 80, 2008.

[15] I. Stamos and P. K. Allen. Geometry and texture recovery of scenes of large scale. *Comput. Vis. Image Underst.*, 88:94–118, 2002.

[16] I. Stamos, L. Liu, C. Chen, G. Wolberg, G. Yu, and S. Zokai. Integrating automated range registration with multi-view geometry for the photorealistic modeling of large-scale scenes. In *IJCV*, volume 78, pages 237–260, 2008.

[17] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *IEEE International Conference on Computer Vision (ICCV)*, pages 839 –846, 1998.

[18] P. H. S. Torr and A. Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78:138–156, 2000.

[19] A. Troccoli and P. Allen. A shadow based method for image to model registration. In *In IEEE Workshop on Image and Video Registration*, pages 169–169, 2004.

[20] R. G. v. Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall. LSD: a line segment detector. *Image Processing On Line*, pages 35–55, 2012.

[21] A. Vasile, F. R. Waugh, D. Greisokh, and R. M. Heinrichs. Automatic alignment of color imagery onto 3d laser radar data. In *Proceedings of the 35th Applied Imagery and Pattern Recognition Workshopn (AIPR)*, page 6, 2006.

[22] L. Wang and U. Neumann. A robust approach for automatic registration of aerial images with untextured aerial lidar data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2623–2630, 2009.

[23] L. Wang, S. You, and U. Neumann. Semiautomatic registration between ground-level panoramas and an ortho-rectified aerial image for building modeling. In *ICCV Workshop on Virtual Representation and Modeling of Large-scale Environments*, pages 8–15, 2007.

[24] G. Yang, J. Becker, and C. Stewart. Estimating the location of a camera with respect to a 3d model. In *3DIM*, pages 159–166, 2007.

[25] W. Zhao, D. Nister, and S. Hsu. Alignment of continuous video onto 3d point clouds. *PAMI*, 27:1305–1318, 2005.