Automatic Registration of LIDAR and Optical Images of Urban Scenes

Andrew Mastin,[1,2], Jeremy Kepner,[2], John Fisher III[1]
[1]Computer Science and Artificial Intelligence Laboratory
[2]Lincoln Laboratory
Massachusetts Institute of Technology, Cambridge MA 02139
mastin@csail.mit.edu, kepner@ll.mit.edu, fisher@csail.mit.edu

Abstract

Fusion of 3D laser radar (LIDAR) imagery and aerial optical imagery is an efficient method for constructing 3D virtual reality models. One difficult aspect of creating such models is registering the optical image with the LIDAR point cloud, which is characterized as a camera pose estimation problem. We propose a novel application of mutual information registration methods, which exploits the statistical dependency in urban scenes of optical apperance with measured LIDAR elevation. We utilize the well known downhill simplex optimization to infer camera pose parameters. We discuss three methods for measuring mutual information between LIDAR imagery and optical imagery. Utilization of OpenGL and graphics hardware in the optimization process yields registration times dramatically lower than previous methods. Using an initial registration comparable to GPS/INS accuracy, we demonstrate the utility of our algorithm with a collection of urban images and present 3D models created with the fused imagery.

1. Introduction

Virtual reality 3D models are useful for understanding a scene of interest. Urban 3D modeling has also gained popularity in entertainment and commercial applications, and has been implemented with geographical image libraries such as Google Earth and Live Search Maps [5, 11]. 3D models are valuable for applications such as urban planning and simulation, interpretation of reconnaissance data, and real-time emergency response. Models are constructed by texture mapping aerial and ground images onto 3D geometry models of the scene. While geometry models have traditionally been constructed manually, recent advances in airborne laser radar (LIDAR) imaging technology have made the acquisition of high resolution digital elevation data more efficient and cost effective.

One challenge in creating such models is registering 2D optical imagery with the 3D LIDAR imagery. This can be formulated as a camera pose estimation problem where the transformation between 3D LIDAR coordinates and 2D image coordinates is characterized by camera parameters such as position, orientation, and focal length. Manual camera pose selection is difficult as it requires simultaneous refinement of numerous camera parameters. Registration can be achieved more efficiently by manually selecting pairs of correspondence points, but this task becomes laborious for situations where many images must be registered to create large 3D models. Some methods have been developed for performing automatic registration, but they suffer from being computationally expensive and/or demonstrating low accuracy rates. In this paper, we discuss a novel methodology for performing automatic camera pose estimation wherein we exploit the observed statistical dependency between LIDAR elevation and optical appearance in urban scenes. We also consider an additional attribute available from some LIDAR devices and investigate its utility for registration.

There has been a considerable amount of research in registering multi-view optical images with LIDAR imagery and other geometric models. Liu, et al. applied structure-from-motion to a collection of photographs to infer a sparse set of 3D points, and then performed 3D-3D registration [10]. While 2D-3D registration is considered, their work emphasizes 3D-3D registration. Zhao, et al. used stereo vision techniques to infer 3D structure from video sequences, followed by 3D-3D registration with the iterative closest point (ICP) algorithm [20]. Both of these methods demonstrate notable results with little or no prior camera information, such as global positioning system (GPS) data, but they require numerous overlapping images of the scene of

This research was partially supported by the Air Force Office of Scientific Research under Award No. FA9550-06-1-0324. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the Air Force. A. Mastin was supported in part by an NDSEG Fellowship.

interest.

In the area of single-view registration, Vasile, et al. used LIDAR data to derive a pseudo-intensity image with shadows for correlation with aerial imagery [17]. Their registration procedure starts with GPS and camera line of sight information and then uses an exhaustive search over translation, scale, and lens distortion. Frueh, et al. developed a similar system based on detection and alignment of line segments in the optical image and projections of line segments from the 3D image [4]. Using a prior camera orientation with accuracy comparable to that of a GPS and inertial navigation system (INS), they used an exhaustive search over camera position, orientation, and focal length. Their system requires approximately 20 hours of computing time on a standard computer. Both methods demonstrate accurate registration results, but are computationally expensive.

There are a variety of algorithms that utilize specific image features to perform registration. Troccoli and Allen used matching of shadows to align images with a 3D model [16]. This requires a strong presence of shadows as well as knowledge of the relative sun position when the photographs were taken. Kurazume, et al. used detection of and matching of edges for registration [7]. One drawback of this method is that it requires a relatively dense 3D point cloud to infer edges. Stamos and Allen used matching of rectangles from building facades for alignment[15]. Yang, et al. use feature matching to align ground images, but they work with a very detailed 3D model [19]. These methods are not robust for all types of urban imagery, and are not optimal for sparse point clouds.

Other approaches have employed vanishing points. Lee, et al. extracted lines from images and 3D models to find vanishing points [8]. Their system cannot register all types of imagery, as it was designed for ground-based images with clearly visible facades. Ding et al. used vanishing points with aerial imagery to detect corners in a similar manner, and used M-estimator sample consensus to identify corner matches [1]. Starting with a GPS/INS prior, their algorithm runs in approximately 3 minutes, but only achieves a 61% accuracy rate for images of a downtown district, a college campus, and a residential region. Liu and Stamos used vanishing points and matching of features to align ground images with 3D range models [9]. All of these approaches are dependent on the strong presence of parallel lines to infer vanishing points, which limits their ability to handle different types of imagery.

2. Methodology

We suggest a straightforward approach that combines an information-theoretic similarity measure with optimization over parameters in a camera model. Our optical images are oblique aerial photographs provided by [14]. The LIDAR data is an ungridded 3D point cloud where each point has an x, y, and z value, as well as probability of detection value that represents how many photons the LIDAR sensor measured. The LIDAR data has a planimetric density of approximately 6 points per square meter. Our problem is a 2D-3D registration problem, but is more accurately characterized as a camera pose estimation problem, where the camera parameters corresponding to the optical image must be estimated. Here we briefly review the criterion used for registration, the camera model, our approach for obtaining ground truth registration, and the method for 3D model generation.

2.1. Mutual Information Registration

Statistical and information-theoretic methods have been used extensively for multi-modal registration of medical imagery. Methods based on these principals have demonstrated excellent performance for a wide variety of 2D-2D and 2D-3D registration applications, e.g. [21, 22]. The methods were originally proposed contemporaneously by Viola and Wells [18] and Maes *et al* [12]. Since their original proposal, these methods (and variations thereof) have become the standard method for automatic registration of dense volumetric medical imagery (e.g. CT and MRI). As these algorithms use grayscale intensity values to evaluate statistics, there is not a direct way to apply them to our problem. Consequently, we attribute features to both types of imagery and evaluate registration statistics over the features.

Since our problem involves registration of imagery in two dimensions with a point cloud in three dimensions, we evaluate registration statistics in the 2D image plane via projection of the LIDAR features within the constraints of a camera model for comparison with the image features. We define $u(x, y)$ and $v(x, y)$ as the the image features and projected LIDAR features on the x-y image plane such that the images are correctly registered. We denote v_o as the initial unregistered projection of LIDAR features obtained from a user selected pose approximation or GPS/INS data. For a specific camera matrix T (defined in Section 2.2), the projected LIDAR features are given by v_T.

Mutual information (MI) based registration methods seek the camera matrix that maximizes the MI between the distribution of photograph features and projected LIDAR features:

$$T_{\mathrm{MI}} = \underset{T}{\operatorname{argmax}}\, I(u; v_T). \tag{1}$$

One definition of mutual information is the Kullback-Leibler (KL) divergence of the joint distribution and the product of marginals:

$$I(u; v_T) = D(p(u, v; T)||p(u)p(v; T)). \tag{2}$$

Consequently, maximizing MI is equivalent to maximizing the KL divergence between the joint distribution under evaluation and the case where the images are independent. This

inherently assumes that the images are best aligned when their statistical dependence is high. KL divergence, defined as an expectation can be approximated from a sample histogram as

$$I(u; v_T) = \int\int p(u, v; T) \log\left(\frac{p(u, v; T)}{p(u)p(v; T)}\right) dudv \quad (3)$$

$$\approx \sum_{i=1}^{N_u}\sum_{j=1}^{N_v} \hat{p}(u_i, v_j; T) \log\left(\frac{\hat{p}(u_i, v_j; T)}{\hat{p}(u_i)\hat{p}(v_j; T)}\right) \quad (4)$$

where $\hat{p}(\bullet)$ denotes a marginal or joint histogram estimate of a density and N_u and N_v denote the number of distinct bins for each modality. See [22] for a detailed discussion of these approximations. All of the features that we work with have 8-bit precision, yielding N_u, N_v = 256. Our images all have dimensions of 1002 x 668 pixels to give approximately 600,000 samples over which to estimate the PMF.

Mutual information can also be expressed in terms of entropies of the LIDAR features, optical features, and their joint entropy:

$$I(u; v_T) = H(u) + H(v_T) - H(u, v_T). \quad (5)$$

In our case, the entropy of the image features remains constant, and the entropy of the LIDAR features remains approximately constant for small perturbations. Accordingly, the calibration matrix that minimizes the joint entropy is a sufficient approximation for the calibration matrix that maximizes the mutual information. That is, minimizing

$$H(u; v_T) \approx \sum_{i=1}^{N_u}\sum_{j=1}^{N_v} \hat{p}(u_i, v_j; T) \log\left(\hat{p}(u_i, v_j; T)\right) \quad (6)$$

is an equivalent to maximizing MI over T.

2.2. Camera Model

The camera model we use is the finite projective camera described in [6]. The transformation from 3D homogeneous coordinates to 2D homogeneous coordinates is given by the 3×4 matrix

$$T = KR[\, I \;\; | -C \,] \quad (7)$$

where $C = [C_x, C_y, C_z]^T$ is the camera center, I is the identity matrix, and R is the rotation matrix describing the orientation of the camera. R is given by the product of rotation matrices

$$R = \begin{bmatrix} c\gamma c\beta & c\alpha s\gamma + s\alpha c\gamma s\beta & s\gamma\alpha - c\gamma c\alpha s\beta \\ -c\beta s\gamma & c\alpha c\gamma - s\alpha s\gamma s\beta & s\alpha c\gamma + c\alpha s\gamma s\beta \\ s\beta & s\alpha c\beta & c\alpha c\beta \end{bmatrix} \quad (8)$$

where α, β, and γ are the Euler angles describing yaw, pitch, and roll. The notation c indicates cosine while s indicates sine. The matrix K is the camera calibration matrix and has the form

$$K = \begin{bmatrix} f_x & t & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{bmatrix} \quad (9)$$

where f_x and f_y are the focal lengths in the x and y directions, (x_0, y_0) are the coordinates of the principal point, and t is the skew. The principal point indicates the location of the center of the image on the image plane and the skew determines the angle between the image plane x- and y-axis. For our images, the skew and principal point parameters are unnecessary, yielding t, x_o, y_o = 0. Additionally, f_x = f_y under the assumption of square pixels. This leaves a constrained finite projective camera parameterized by seven variables: C_x, C_y, C_z, α, β, γ, and the field-of-view (which can be computed from the focal length).

2.3. Ground Truth Registration

Expert chosen correspondence points are utilized to determine ground truth. Correspondence points are chosen by identifying salient geometric features visible in both images. We use the algorithm given by [6, p. 184] to determine the best camera matrix. That algorithm minimizes the sum of squared algebraic errors between correspondence points in the image plane,

$$T_G = \underset{T}{\operatorname{argmin}} \sum_i d_{alg}(\mathbf{X}'_i, T\mathbf{X}_i) \quad (10)$$

where $\mathbf{X}'$ represents a 2D homogeneous point and $\mathbf{X}$ represents a 3D homogeneous point. All images are registered using 30 correspondence points.

2.4. Registration Algorithm

Our algorithm renders 3D points that are projected onto the image plane for evaluating statistics. The point cloud is rendered in OpenGL for each iteration. The LIDAR point cloud data sets tend to be very large (on the order of millions of points). Newer graphics cards have sufficient memory to store the entire data set in graphics card memory, which makes 3D rendering extremely efficient. For all MI measures tested, we used downhill simplex optimization, which is derivative free [13].

An important issue is how to attribute the 3-D LIDAR point cloud data so that information-theoretic methods can be applied. We use three different methods for evaluating mutual information between the two image modes. The

first is simply the mutual information between elevation in the LIDAR point cloud and luminance in the optical image. The point cloud is rendered with height intensities, where brighter points indicate a higher elevation. Only image pixels that have a corresponding projected LIDAR point are used for calculating registration statistics. While this is a simple attribution, for urban scenes it provides useful results (as we will demonstrate). The intuition for this simple feature is that the visual appearance of urban scenes tend to vary in a structured way by height for architectural reasons. Consequently, there is measurable dependence between the optical appearance and the measured LIDAR height. A scene shown by both modalities is shown in Figure 1(a) and (b). We see that similar structure exists across both images.

The second measure that we use is the mutual information between the luminance in the optical image and probability of detection (pdet) values in the LIDAR point cloud. Probability of detection values indicate the number of photons returned to the LIDAR sensor for each point, yielding a point cloud representation that looks similar to a grayscale aerial image, as shown in Figure 1(c). For example, one can see details of the walkway between buildings.

Finally, we consider the joint entropy among optical image luminance, LIDAR elevation, and LIDAR pdet values. We approximate the LIDAR pdet values as statistically independent of the elevation values conditioned on the optical image luminance values. This leads to the following approximation of joint entropy:

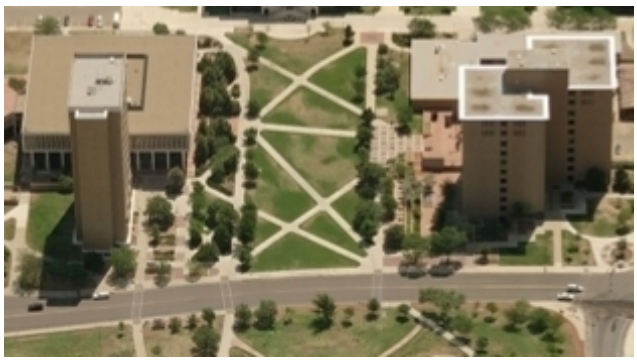
$$H(u, v_e, v_p) = H(u, v_e) + H(u, v_p), \qquad (11)$$

where u is the image luminance, v_e is the LIDAR elevation, and v_p is the LIDAR pdet values.

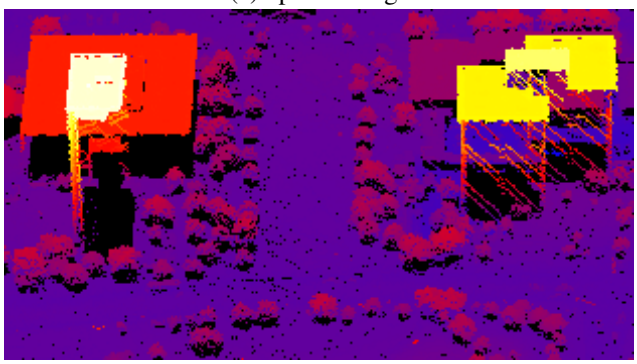
2.5. 3D Model Generation

We create 3D models by texture mapping registered optical images onto a mesh that is inferred from the LIDAR point cloud. Since the sampling rate of our LIDAR data is relatively high compared to the sizes buildings, performing a Delaunay triangulation produces a mesh that closely represents the true 3D structure. An example of a mesh structure is shown in Figure 2 (a).

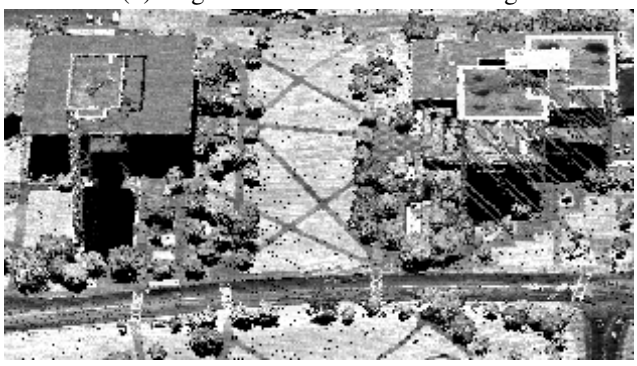
While the resulting mesh without texture has a somewhat jagged appearance, the texture mapped rendering in Figure 2 (b) is relatively smooth. The texture mapping is performed using hardware accelerated shadow mapping to provide occlusion reasoning and automatic texture generation in OpenGL [2, 3]. The dark regions in the figure result from occlusions in the projective texture mapping.



(a) optical image



(b) height-encoded LIDAR rendering

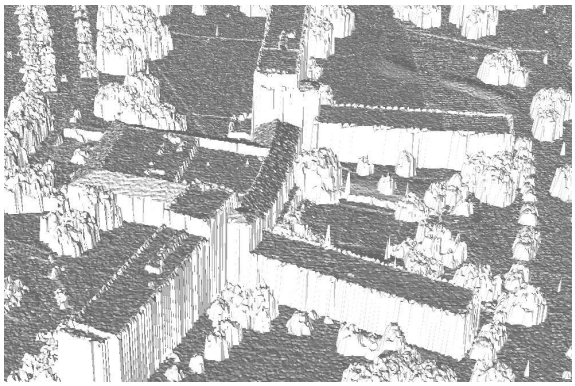


(c) Pdet LIDAR

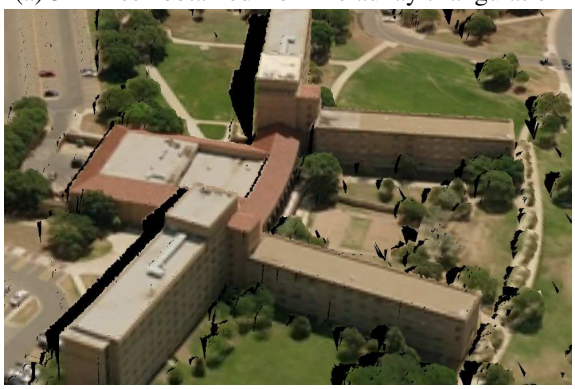
Figure 1. Detail of LIDAR/optical scene. (a) Shows an optical image of two buildings with pathways between, (b) shows the registered LIDAR data set of the same scene with intensity encoded by height, while (c) shows the LIDAR date with the Pdet attribute.

3. Results

We show the results of our algorithm with a collection of eight urban images of Lubbock, Texas. As with previous work, it is conventional to start with an approximate initial registration that is available from a GPS/INS system. We simulate a variety of initial approximate registrations and characterize the accuracy of the final registration using the three different measures of mutual information. Overall, the results demonstrate that the MI measures are reliable for registration and that the algorithm is fast.



(a) 3-D mesh obtained from Delaunay triangulation



(b) Registered image textured onto 3-D mesh

Figure 2. For dense LIDAR measurements, Delaunay triangulation provides sufficient structure for registration of optical images onto a 3D model. The image at left shows a detail of the mesh over a larger area, while the image at right shows the resulting texture map after registration.

For each of the eight images, we simulated 100 initial coarse registrations. Table 1 describes the range of camera parameter perturbations over which the registration method was tested. Some of the angle perturbations seem small, but it is important to note that the long standoff distance magnifies the effect of the Euler angles and field-of-view angle. For example, Figure 3 shows a blend of the LIDAR data and the corresponding image with an α perturbation of one degree. Our images did not contain focal length information, so we included the field-of-view parameter in our parameter search. When the focal length is known, the complexity of the search is reduced to six parameters. We randomly sampled parameters from this range to obtain initial parameters for registration.

It was apparent by simple inspection whether or not the images were correctly registered. In all cases, the image was either registered close enough for projective texture mapping or clearly unaligned. Examples of post-registration texture maps are shown in Figure 2 (b) and Figure 5 (b).

Results of the experiments are shown in Tables 2,3, and 4, describing the number of successes, duration of the optimization, and the number of iterations, respectively. As expected, the dual measure of MI (which uses LIDAR elevation and pdet) demonstrates the best results, with an overall accuracy of 98.5%. However, it is interesting to note that marginal benefit of using the pdet MI measure (95.8%) and the dual MI measure (98.5%) over the elevation MI measure (93.5%) is relatively small. This an important result since not all LIDAR sensors provide pdet values. Table 3 shows fast registration times, which all averaged to be less than 20 seconds. The dual registration times are approximately the sum of the elevation and pdet registration times, which is expected since two images must be rendered for each iteration with our implementation. An example of an initial and final alignment are shown in Figure 6.

Figure 4 depicts the measured joint entropy for an image as camera parameters are varied smoothly from the ground truth. While the camera parameter being evaluated is varied, the other parameters are held constant at their correct values. The plots demonstrate adequate quasiconvexity, which is necessary for downhill simplex optimization. All seven parameters exhibit similar behavior; the parameters plotted yield the smallest capture range, and so, in some sense, reflect worst-case results.

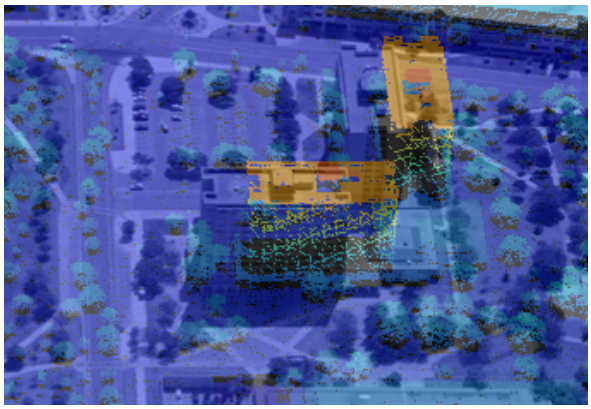


Figure 3. Fade between LIDAR image and optical image with an α perturbation of one degree.

Parameter	Range	Units
C_x	20	meters
C_y	20	meters
C_z	20	meters
α	0.5	degrees
β	0.5	degrees
γ	5	degrees
fov	0.5	degrees

Table 1. Range of camera parameter perturbations.

JE Measure:	Elev	Pdet	Dual
Image 1	100	93	100
Image 2	97	99	99
Image 3	83	100	96
Image 4	76	83	97
Image 5	97	97	99
Image 6	100	100	100
Image 7	96	95	97
Image 8	99	99	100
Total:	93.5%	95.8%	98.5%

Table 2. Number of correctly registered images (out of 100) randomly sampled perturbations.

JE Measure:	Elev	Pdet	Dual
Image 1	6.30	8.12	14.18
Image 2	6.80	7.14	13.39
Image 3	6.44	8.59	14.67
Image 4	5.76	5.79	10.88
Image 5	6.55	6.58	12.39
Image 6	6.19	7.13	11.89
Image 7	6.77	8.27	14.44
Image 8	5.99	6.53	12.40
Total:	6.35	7.27	13.03

Table 3. Mean registration times in seconds (for correctly registered images)

JE Measure:	Elev	Pdet	Dual
Image 1	114.8	115.9	112.1
Image 2	127.4	123.1	118.5
Image 3	116.7	119.9	115.0
Image 4	98.5	104.0	99.6
Image 5	112.4	116.6	112.9
Image 6	114.9	116.9	110.0
Image 7	118.4	115.2	110.5
Image 8	111.2	118.4	108.8
Total:	115.6	116.2	110.9

Table 4. Mean number of iterations (for correctly registered images)

4. Model Renderings

As described previously, 3-D mesh structures are constructed over the entire LIDAR point cloud using Delaunay triangulation. Figures 5 depicts optical images of the same area from three perspectives. Next to these are the rendered 3D models from an off-camera view. All of these images were registered using the automatic registration method described.

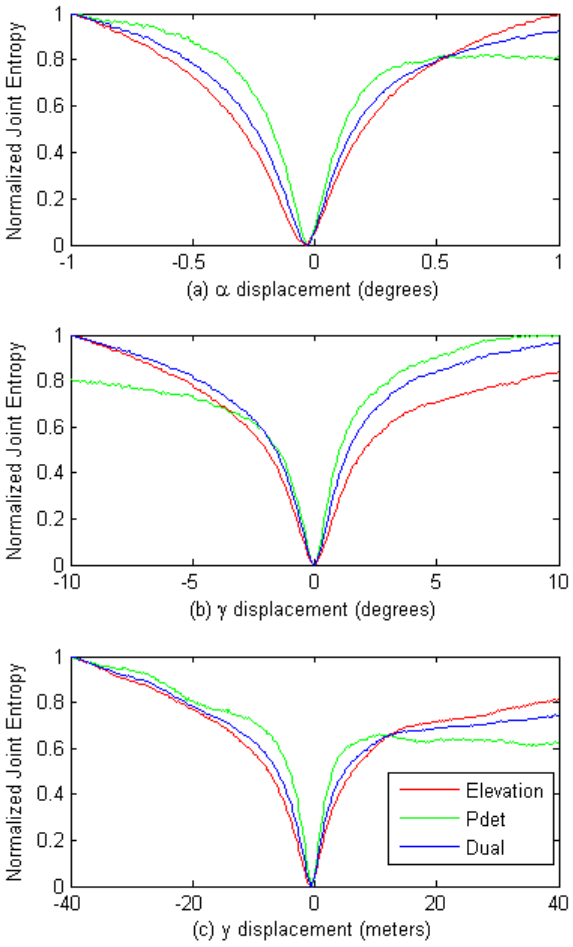
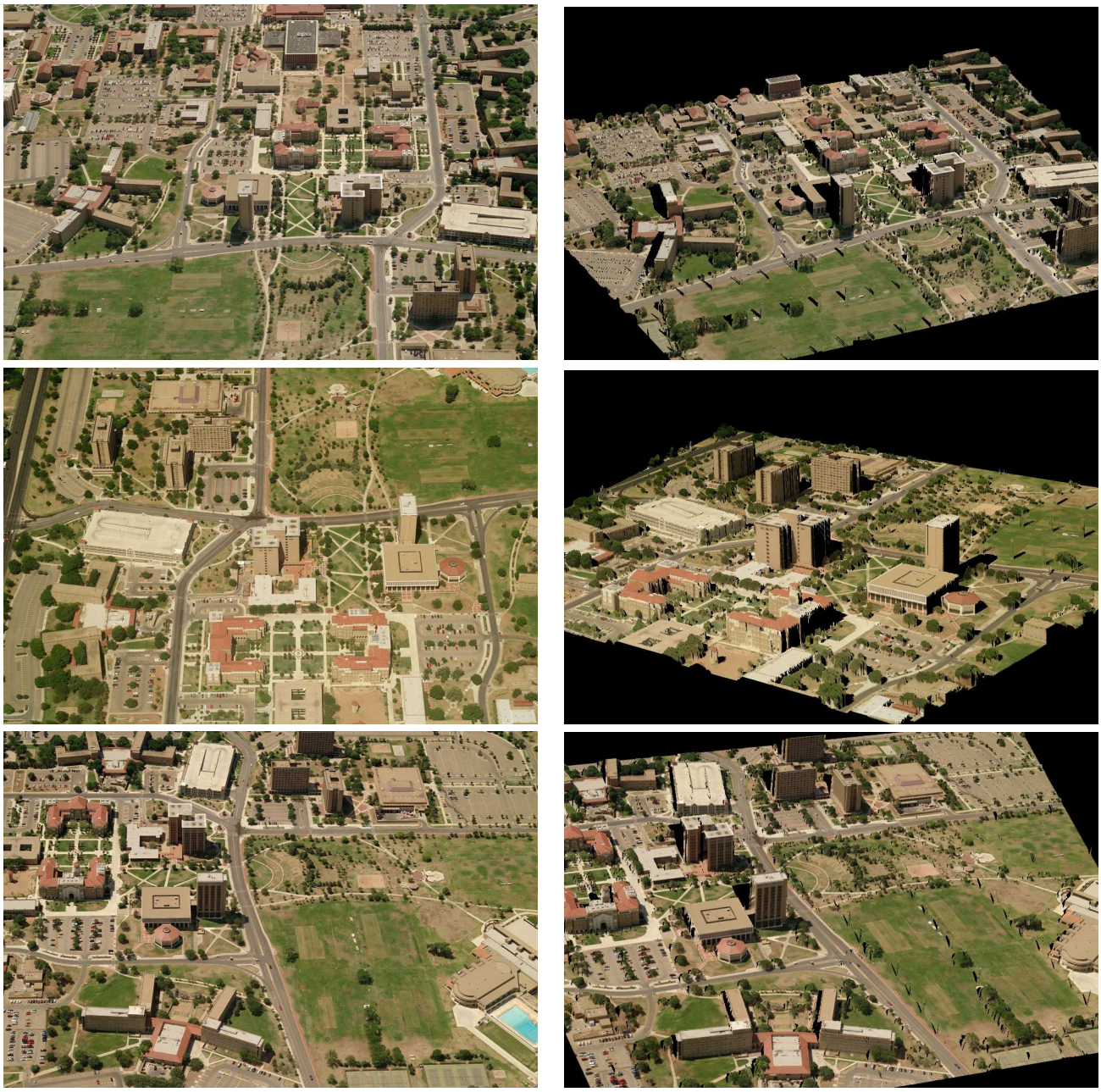


Figure 4. Plots of normalized joint entropy as three of the camera parameters are smoothly perturbed from the point of registration.

5. Conclusions

We have presented a novel application of mutual information for registration of 3D laser radar (LIDAR) imagery and aerial optical imagery. The approach is efficient in that it is fast and automatic for constructing 3D virtual reality models of urban scenes. Registration is achieved using 3D-2D renderings of height and probability of detection attributes of the LIDAR. Empirical results demonstrate that these attribute choices are suitable for achieving dense registration over these modalities, with registration accuracies averaging over 90%. The results also show that having only elevation information for the LIDAR data provides notable registration results, while using pdet values provides a slight benefit in registration accuracy. We further characterized the robustness of the approach via probing experiments in



(a) optical images (b) texture maps after registration

Figure 5. The column on the left depicts optical images of the same scene from three different perspectives. Results of texture mapping post-registration are shown on the right.

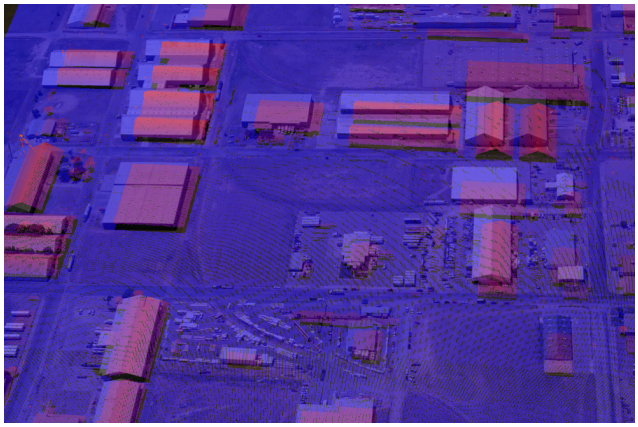
which we randomly perturbed the camera parameters from known ground truth. Our results show the bounds of camera parameter perturbations over which a reliable registration can be achieved; on the order of 20 meters of displacement; 0.5 degrees of yaw, pitch, and field-of-view; and 5 degrees of roll. The method was implemented in OpenGL utilizing advanced graphics hardware in the optimization process, yielding registration times on the order of seconds.

References

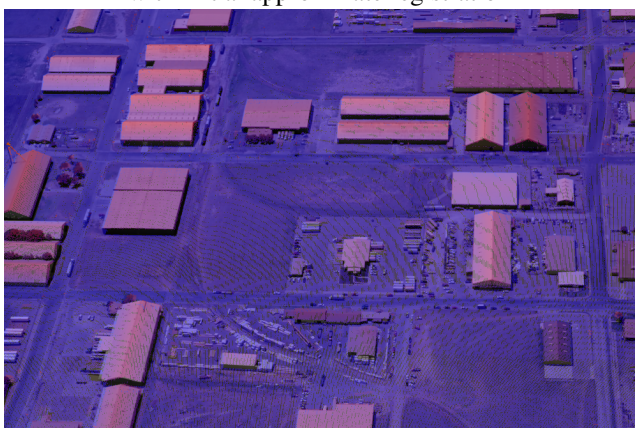
[1] M. Ding, K. Lyngbaek, and A. Zakhor. Automatic registration of aerial imagery with untextured 3d lidar models. *CVPR*, 2008. 2

[2] C. Everitt. Projective texture mapping. Technical report, NVIDIA, http://developer.nvidia.com/, 2001. 4

[3] C. Everitt, A. Rege, and C. Cebenoyan. Hardware shadow mapping. Technical report, NVIDIA,



(a) Fade between LIDAR image and optical image with initial approximate registration



(b) Fade between LIDAR image and optical image with post-algorithm registration

Figure 6. Example of registration results with image superimposed on projected LIDAR image

http://developer.nvidia.com/, 2000. 4

[4] C. Frueh, R. Sammon, and A. Zakhor. Automated texture mapping of 3d city models with oblique aerial imagery. *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on*, pages 396–403, 6-9 Sept. 2004. 2

[5] Google Earth, 2008. http://earth.google.com/. 1

[6] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000. 3

[7] R. Kurazume, K. Nishino, M. D. Wheeler, and K. Ikeuchi. Mapping textures on 3d geometric model using reflectance image. *Syst. Comput. Japan*, 36(13):92–101, 2005. 2

[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(3), 2002. 2

[9] L. Liu and I. Stamos. A systematic approach for 2d-image to 3d-range registration in urban environments. pages 1–8, Oct. 2007. 2

[10] L. Liu, I. Stamos, G. Yu, G. Wolberg, and S. Zokai. Multiview geometry for texture mapping 2d images onto 3d range data. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2293–2300, Washington, DC, USA, 2006. IEEE Computer Society. 1

[11] Live Search Maps, 2008. http://maps.live.com/. 1

[12] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *Medical Imaging, IEEE Transactions on*, 16(2):187–198, April 1997. 2

[13] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, January 1965. 3

[14] Pictometry International, 2008. http://www.pictometry.com. 2

[15] I. Stamos and P. K. Allen. Geometry and texture recovery of scenes of large scale. *Comput. Vis. Image Underst.*, 88(2):94–118, 2002. 2

[16] A. Troccoli and P. Allen. A shadow based method for image to model registration. pages 169–169, June 2004. 2

[17] A. Vasile, F. R. Waugh, D. Greisokh, and R. M. Heinrichs. Automatic alignment of color imagery onto 3d laser radar data. In *AIPR '06: Proceedings of the 35th Applied Imagery and Pattern Recognition Workshop*, page 6, Washington, DC, USA, 2006. IEEE Computer Society. 2

[18] P. Viola and W. Wells III. Alignment by maximization of mutual information. In *Proceedings of IEEE International Conference on Computer Vision*, pages 16–23, 1995. 2

[19] G. Yang, J. Becker, and C. Stewart. Estimating the location of a camera with respect to a 3d model. pages 159–166, Aug. 2007. 2

[20] W. Zhao, D. Nister, and S. Hsu. Alignment of continuous video onto 3d point clouds. *CVPR*, 02:964–971, 2004. 1

[21] L. Zöllei, E. Grimson, A. Norbash, and W. Wells. 2d-3d rigid registration of x-ray fluoroscopy and ct images using mutual information and sparsely sampled histogram estimators. *CVPR*, 2:696, 2001. 2

[22] L. Zöllei, J. W. F. III, and W. M. W. III. A unified statistical and information theoretic framework for multi-modal image registration. In *IPMI*, pages 366–377, 2003. 2, 3