FUSING STRUCTURE FROM MOTION AND LIDAR FOR DENSE ACCURATE DEPTH MAP ESTIMATION

Li Ding and Gaurav Sharma

Dept. of Electrical and Computer Engineering, University of Rochester, Rochester, NY

ABSTRACT

We present a novel framework for precisely estimating dense depth maps by combining 3D lidar scans with a set of uncalibrated camera RGB color images for the same scene. Rough estimates for 3D structure obtained using structure from motion (SfM) on the uncalibrated images are first co-registered with the lidar scan and then a precise alignment between the datasets is estimated by identifying correspondences between the captured images and reprojected images for individual cameras from the 3D lidar point clouds. The precise alignment is used to update both the camera geometry parameters for the images and the individual camera radial distortion estimates, thereby providing a 3D-to-2D transformation that accurately maps the 3D lidar scan onto the 2D image planes. The 3D to 2D map is then utilized to estimate a dense depth map for each image. Experimental results on two datasets that include independently acquired high-resolution color images and 3D point cloud datasets indicate the utility of the framework. The proposed approach offers significant improvements on results obtained with SfM alone.

Index Terms— structure from motion, lidar, depth map, sensor fusion

1. INTRODUCTION

Contemporary techniques for estimating structure from motion (SfM) allow us to conveniently exploit multiple images of a scene, captured from different viewpoints using only consumer-grade cameras, to jointly estimate both the 3D structure of the scene and the parameters of the cameras used for the captured images [1, 2]. Although these techniques are powerful, they have limitations: 3D structure is directly estimated for only a sparse set of points for which correspondences can be reliably established between the multiple viewpoints and the accuracy of 3D locations of the points as well as the camera parameters is limited because of the deviations that actual consumer cameras exhibit compared with the ideal imaging models used in SfM. Other than the sparse set of corresponding points between different viewpoints, 3D structure is obtained by interpolation, which can be problematic, particularly for untextured regions in the scene. New sensing modalities, such as a lidar and structured-light based depth sensing, provide an alternative approach for sensing 3D structure. These modalities allow for precise estimation of relatively dense 3D structure and work well even in untextured regions. However, data capture for these modalities is considerably more involved than for simple point and shoot cameras and there is either no accompanying color imagery or such imagery is available only at low resolution.

In this paper, we propose a novel methodology for synergistically fusing 3D structure from lidar with SfM to obtain a more accurate and higher resolution 3D representation than is achieved with each modality alone. Specifically, we demonstrate our methodology by obtaining a high resolution precise depth map for each of the images. The proposed method has utility in applications where

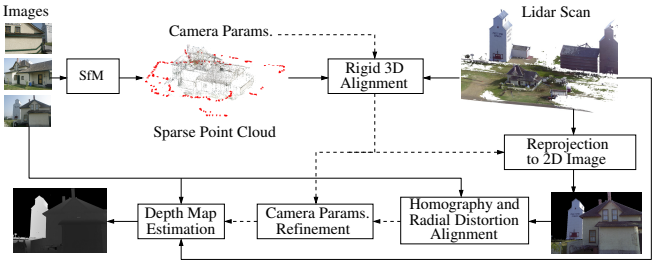


Fig. 1: Overview of the proposed methodology for fusing structure from motion (SfM) and lidar scan. The solid and dashed lines show the data flow and parameters flow, respectively.

a combination of high resolution geometry and texture (color) is desired for instance in improved photorealistic rendering of urban scenes [3, 4], augmented reality [5], self-driving vehicles [6] and cultural heritage preservation [7, 8].

Fusing the data from the different sensors relies on the registration between 3D lidar scan and 2D photographs. Several methods have been proposed for this problem. Traditional approaches are based on feature matching where a set of points or lines correspondences are known. Zhang and Pless [9] use a checkerboard as target for point matching that is observed by camera and laser scanner. The camera is pre-calibrated in their set-up and only the extrinsic parameters are estimated. In [3], 3D lines are extracted from the lidar scan and matched with 2D features generated from vanishing points. Ding et al [4] apply a two-step algorithm that first recovers a coarse camera parameter and then matches the 2D corners from image with the orthogonal 3D corners. Tamas and Kato [10] use shape registration that uses planar objects visible in both sensors. More recent work [11, 12] exploits information-theoretic methods for this problem. In [11], the mutual information between reflectivity recorded by lidar and pixel intensity in image is maximized to determine camera pose. Mastin et al [12], on the other hand, formulate the registration problem as maximizing the mutual information between the distribution of point features in the image and lidar data.

Several algorithms integrate multi-view geometry to reconstruct a point cloud from a set of unordered images [13, 8] or video sequence [14]. The basic idea of these algorithms is to reconstruct a sparse point cloud and to align with lidar scan. In [14], the problem of point cloud alignment is solved by the ICP algorithm, which has limited ability for dealing with outliers. Liu et al [13] determine camera extrinsic parameters based on a set of 3D corresponding points. However, in their methodology the images have to be pre-calibrated in order to achieve satisfactory accuracy. Subsequent work of [8] proposes a similar framework for 3D-to-2D registration, requiring human interaction to align several images to the lidar scan by manually determining the point correspondences.

Other prior works include: in [15], a joint calibration and sensor fusion algorithm is applied by alignming the edges in depth map and

intensity image. In [16], first a region matching method is used to match the lidar with an image from the similar viewpoint, and the all images are aligned together to handle the changes in viewpoint.

In this paper, we introduce a novel framework for fusing the different data modalities for 3D lidar scans and 2D photography to estimate dense depth maps. In contrast with the work in [17] where the lidar and camera sensors are rigidly attached on a vehicle, our approach does not require the relative position of lidar and camera to be fixed and the sensors can conveniently be deployed independently for data acquisition. Additionally, instead of determining a set of cross-domain feature correspondences in 3D space and 2D image plane, which is error-prone, we combine SfM with lidar scan and obtain an accurate transformation by matching images in 2D reprojection space. Another benefit of the 3D-to-2D alignment method we use is that both intrinsic and extrinsic camera parameters can be estimated, whereas, when 3D to 3D alignment [9, 15] the relative change in geometry between the different sensors can only provide extrinsic parameters, which are not precise enough for high resolution imagery. The proposed automated framework does not require prior camera calibration and is robust to outliers.

The paper is organized as follows. Section 2 describes a sketch of the proposed framework. We present the experimental results on two real datasets in Section 3, and conclude the paper in Section 4.

2. FUSING SFM AND LIDAR

The proposed framework addresses the problem of estimating depth maps from 3D lidar scans and a set of 2D images using the pipeline depicted in Fig 1. The lidar scan and the images form the input and in order to accomplish the fusion of these modalities, we obtain precise estimates of both the camera intrinsic parameters and the geometric parameters that relate different sensor coordinate systems, including radial distortion, which is critical for high resolution imagery. To accomplish our goal, we apply a two-stage process to automatically recover an accurate transformation that maps 3D lidar scan onto 2D image plane. This transformation is then utilized to estimate depth map of the corresponding 2D image. In the first stage, a sparse point cloud is reconstructed by incremental SfM algorithm and is aligned with lidar scan to obtain initial camera parameters. In the second stage, we align each input image with a corresponding synthetic image that is generated from lidar scan at the same viewpoint, where radial distortion is considered to refine the initial camera parameters.

2.1. Initial 3D-to-2D Transformation Estimation

The first stage aims to estimate coarse camera parameters with respect to lidar scan. The input to this stage comprises a point cloud M_l captured by lidar and a series of uncalibrated images $\mathbf{I} = \{I_n\}_{n=1}^{N}$ observing the scene, where N is the number of images. Each point in the point cloud is associated with 3D coordinate and color data, and is denoted by

$$M_l = \{\left(B_i^l, V_i^l\right)\}_{i=1}^{K_l}, \tag{1}$$

where $B_i^l = \left(X_i^l, Y_i^l, Z_i^l\right)^T$ is the 3D coordinate of the point i, $V_i^l = (R_i, G_i, B_i)^T$ is the RGB color value for the point recorded by lidar, and K_l is the number of points in the point cloud. This stage returns initial camera parameters that map the 3D point onto each image plane of the camera. The camera model is given by [1]

$$\tilde{b}_i = \mathbf{K}_n \left[\mathbf{R}_n | \mathbf{t}_n\right] \tilde{B}_i, \tag{2}$$

where $\tilde{b}_i = (x_i, y_i, 1)^T$ denotes the homogeneous coordinate on the 2D image plane of the projected 3D point $\tilde{B}_i = (X_i, Y_i, Z_i, 1)^T$, the extrinsic matrix, $[\mathbf{R}_n|\mathbf{t}_n]$, contains a rotation matrix $\mathbf{R}_n$ and a translation vector $\mathbf{t}_n$ of the camera, and $\mathbf{K}_n$ is the intrinsic matrix

$$\mathbf{K}_n = \begin{bmatrix} \alpha_n & \gamma_i & u_n \\ 0 & \beta_n & v_n \\ 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

where α_n and β_n are the focal length in the horizontal and vertical direction of the camera, respectively, (u_n, v_n) is the principal point, and γ_i denotes for the skew parameter, which is equal to 0 for most cameras. The camera parameters, which describes the projection of a 3D point onto 2D image plane, is given by $\mathbf{P}_n = \mathbf{K}_n \left[\mathbf{R}_n|\mathbf{t}_n\right]$.

Our approach to recover initial camera parameter is based on the SfM technique, which simultaneously reconstructs 3D structure and camera positions and orientations from a set of images captured at various viewpoints. Among several proposed SfM strategies, incremental SfM [18, 2, 19] has been widely used. The basic idea is that a set of keypoints are detected in each image and matched between all pairs of images. Then an iterative procedure is performed to recover camera parameters as well as 3D scene. In each iteration, only one camera is added for optimization.

We use a similar notation to represent the sparse point cloud computed by SfM

$$M_s = \{\left(B_j^s, V_j^s\right)\}_{j=1}^{K_s}, \tag{4}$$

where $B_j^s = \left(X_j^s, Y_j^s, Z_j^s\right)^T$ is the coordinate of point j, and $V_j^s = (R_j^s, G_j^s, B_j^s)^T$ is the color value for the point, and K_S is the number of points in the reconstructed point cloud. SfM also returns a set of camera parameters $\mathbf{P}_n^s = \mathbf{K}_n^s \left[\mathbf{R}_n^s|\mathbf{t}_n^s\right], n = 1, 2, \ldots, N$, with respect to M_s, for the reconstructed cameras.

Given that two point clouds, M_l from lidar and M_s from SfM, are associated with two different coordinate systems, it is necessary to first map them into a common reference by point cloud alignment. We adopt a rigid transformation including a rotation matrix $\mathbf{R}_{3\times3}$, a translation vector $\mathbf{t}_{3\times1}$, and a scaling factor s, and map the points of M_s to M_l. In practice, the sparse-to-dense point cloud alignment is complicated due to different number of points in two point clouds. We adopt the coherent point drift (CPD) algorithm [20] where the transformation is estimated within EM framework [21] that is robust to noise and accommodates different number of points as well. For each reconstructed camera n, the new extrinsic matrix $\left[\mathbf{R}_n^l|\mathbf{t}_n^l\right]$, which relates the camera coordinate system and lidar coordinate system, is computed based on the transformation

$$\begin{aligned} \mathbf{R}_n^l &= s\mathbf{R}_{3\times3}\mathbf{R}_n^s \\ \mathbf{t}_n^l &= s\mathbf{R}_{3\times3}\mathbf{t}_n^s + \mathbf{t}_{3\times1}. \end{aligned} \tag{5}$$

The new camera parameters that provide initial 3D-to-2D transformation between the lidar scan M_l and the input images are $\mathbf{P}_n^l = \mathbf{K}_n^s \left[\mathbf{R}_n^l|\mathbf{t}_n^l\right], n = 1, 2, \ldots, N$.

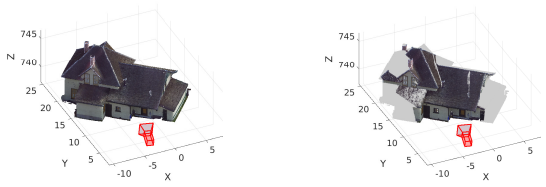
2.2. 3D-to-2D Transformation Refinement

While SfM is a prevalent technology for recovering 3D structure along with camera motions, the state of the art algorithm is still far from producing satisfactory results in terms of accuracy [22, 23] for several reasons. First, SfM applies bundle adjustment as a global refinement, which could only end up in a local minima. Second, the camera model in the first stage does not take into account the radial

distortion, which introduces noticeable error into camera parameters. Therefore, the initial camera parameters $\mathbf{P}_n^l$ are refined in this stage to obtain accurate transformation for estimating a dense depth map.

Our method is motivated by the observation that if we set up a virtual camera using the camera parameter $\mathbf{P}_n^l$, we can synthesize an image $I_n^{'}$ observing the 3D scene M_l and align $I_n^{'}$ with the real image I_n. The registration of this image pair, together with $\mathbf{P}_n^l$, is able to offer a more accurate transformation between M_l and I_n.

One challenging problem we found empirically is the extraction of visible points in M_l from a given camera viewpoint. A single point can not be occluded by another point unless one point lies exactly on the ray from the camera center to another. However, if the image I_n is taken in front of a building, for example, only a subset of points representing the facade is visible, as shown in Fig. 2. Hence, We apply the "hidden" point removal operator [24] to determine whether the point B_i^l is visible from the given camera position $\mathbf{C}_n^l = -\mathbf{R}_n^{l\,-1}\mathbf{t}$, and use the visible points to synthesize the image.



(a) (b)

Fig. 2: Example of points visibility. (a) shows the point cloud and the camera position and orientation. (b) shows the visible points from this camera. The hidden points are shown in gray.

To determine an accurate transformation between two images I_n and $I_n^{'}$, we consider the homography as well as radial distortion. The homography H that relates the coordinates of the corresponding image points (without radial distortion) in two images is

$$s\tilde{\mathbf{x}}_i^{'} = H\tilde{\mathbf{x}}_i, \tag{6}$$

where $\tilde{\mathbf{x}}_i = \left(x_i^{'}, y_i^{'}, 1\right)^T$ and $\tilde{\mathbf{x}}_i^{'} = (x_i, y_i, 1)^T$ denote the correspondence points in I_n and $I_n^{'}$, respectively, H is a nonsingular 3×3 matrix, and s is an arbitrary scale factor.

Radial distortion has to be considered here because the linear camera model in (2) does not directly hold for most cameras. In this paper, we assume that each image is affected by the first-order radial distortion, and use the following model to remove the distortion [1]

$$L\left(\tilde{\mathbf{x}}_d; k\right) = \left(\frac{x_d}{1+kr^2}, \frac{y_d}{1+kr^2}, 1\right)^T, \tag{7}$$

where $\tilde{\mathbf{x}}_d = (x_d, y_d, 1)^T$ is the coordinate of distorted image point, $L\left(\mathbf{x}_d; k\right)$ is the the distortion-free image point coordinates that obeys the camera model in (2), k is coefficient of distortion, and r is the radial distance $r = \sqrt{x_d^2 + y_d^2}$. The transformation between two corresponding points in the presence of radial distortion is

$$L\left(\tilde{\mathbf{x}}_i^{'}; k_2\right) = HL\left(\tilde{\mathbf{x}}_i; k_1\right), \tag{8}$$

where k_1 and k_2 are the distortion coefficient in I_n and $I_n^{'}$, respectively. We apply the generalized dual bootstrap-ICP algorithm [25] for aligning the pair of images, because it is robust to the differences in viewpoint and able to estimate the distortion in two images.

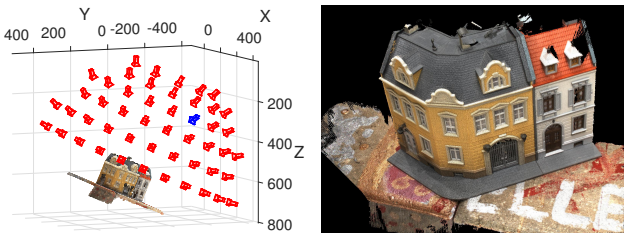
2.3. Depth Map Estimation

By combining the initial camera parameter $\mathbf{P}_n^l$ with the alignment between two images I_n and $I_n^{'}$, we obtain an accurate transformation between the 3D point cloud M_l and the 2D image I_n. The depth information derived from lidar scan is then fused with images to estimate depth map. Notice that the resolution of actual image is much higher than that of depth map. Hence, we perform a post-processing step to enhance the low-resolution range map using a bilateral filter [26]. The process yields a depth map of higher resolution than is feasible with the lidar modality alone.

3. EXPERIMENTAL RESULTS

The proposed framework is extensively tested on two datasets with different characteristics. The first dataset is DTU robot image dataset [22], which contains a point cloud and a set of calibrated images of indoor scenes. The second dataset we used is from the Architectural Biometrics project [27, 28], which contains a 3D point cloud and a set of uncalibrated images of outdoor scene representing a Canadian railway station called *Meeting Creek*.

We first evaluate the accuracy of our framework on the DTU dataset. The camera calibration parameters for each image are provided which can be considered as ground truth to measure the reprojection error. We apply the state of the art SfM algorithm *VisualSFM* [19] to reconstruct the 3D scene from 49 images of resolution 1200×1600, and align it with the dense point cloud. Figure 3 shows the camera positions and orientations with respective to the dense point cloud $\mathbf{P}_n^l$ and an example of synthetic image $I_{17}^{'}$.



(a) (b)

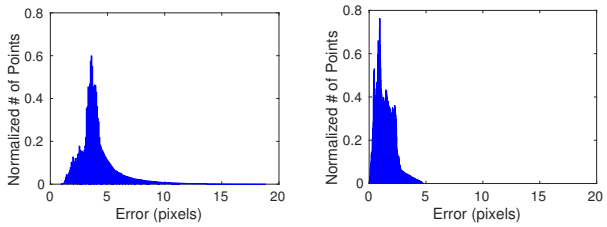
Fig. 3: Results of initial 3D-to-2D transformation estimation using DTU dataset. (a) shows the reconstructed camera with respective to the dense point cloud. (b) shows the synthetic image $I_{17}^{'}$ that corresponds to the blue camera in (a).

After the transformation refinement in the second stage, we calculate the reprojection error and compare it with the result obtained with *VisualSFM* alone. Figure 4 shows the histogram of the reprojection error in pixel units for image 17 in the dataset. The average value of reprojection error, in this image, is reduced from 4.13 pixels (*VisualSFM*) to 1.43 pixels (proposed). We also calculate the 98th percentile of reprojection error that can be viewed as the worst case of 3D-to-2D transformation, which, in this image, are 9.65 pixels from *VisualSFM* and 3.64 pixels from the proposed method.

Figure 5 shows average (top) and 98th percentile (bottom) reprojection error for each image. The magenta and blue bars indicate the results from *VisualSFM* and our method, respectively. We improve the average reprojection error over all images in the dataset by 1.75 pixels, and the 98th percentile error by 4.55 pixels.

Next, we validate the proposed method with the second dataset *Meeting Creek*, which is shown is Fig. 1. The major difference be-

Fig. 6: Sample results for dense depth map estimation. The first row shows five input RGB images captured using a digital camera, and the second row shows the corresponding depth map generated by our method. The corresponding point cloud is shown in the example of Fig. 1.



(a) (b)

Fig. 4: Histogram of the reprojection error using (a) *VisualSFM* and (b) our method for image 17 of the DTU dataset. Both the mean and the 98th percentile of the reprojection error are significantly reduced using our method.

tween this dataset and DTU dataset is that *Meeting Creek* contains a wide outdoor scene including several large buildings and trees, while the objects in DTU dataset are indoor scenes and obtained under a controlled environment. In this experiment, 459 images with resolution 3888×2592 are acquired using the Canon EOS XS camera. Since we do not have any prior information regarding camera calibration, we test the performance of our method by visually evaluating the quality of depth maps. Figure 7 shows an example of comparison between reprojected points using *VisualSFM* and our method. We can readily see the displacement of alignment using *VisualSFM* in Fig. 7a. Our method, however, is able to align the edge points of roof and chimneys sufficiently close to the corresponding image pixels, as shown in Fig 7b. Figure 6 provides the final results for depth map estimation. Each column shows a pair of RGB image and the corresponding depth map. Visual evaluation indicates that our method is able to create accurate dense depth map. It is also worth pointing out several issues in the final depth map shown in Fig 6. First, the window region in the third depth map does not have depth information, which is due to the lack of data in the corresponding area for point cloud. We also notice that, in the fourth depth map, the tree region is visually seen not to be accurate enough. The error can occur either because of the motion of the tree during the scanning or the process of hidden point removal.

4. CONCLUSION

The framework we present in this paper provides an accurate methodology for estimating dense depth maps. Our approach is based on fusing structure from motion and lidar to precisely recover

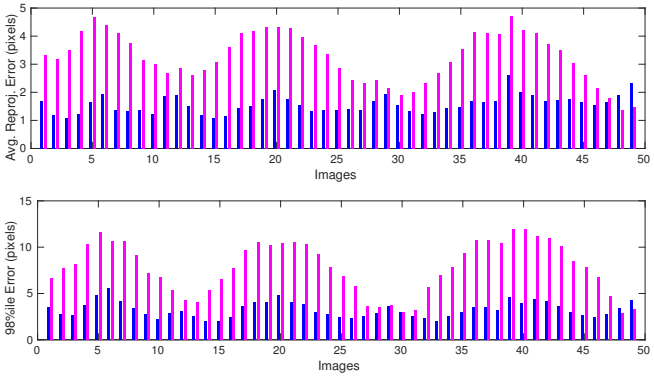


Fig. 5: Reprojection error for each image in DTU dataset. The magenta bars and blue bars indicate the results from *VisualSFM* and our method, respectively. Top: average error for each image; Bottom: the 98 percentile of error for each image.

the transformation from 3D lidar space to 2D image plane. Experimental results on two datasets demonstrate that the framework achieves high accuracy of transformation in terms of reprojection error and generates dense depth maps corresponding to input RGB images.

(a) (b)

Fig. 7: Sample results for visual comparison of reprojecting points onto image planes between (a) *VisualSFM* and (b) our method. The image shows the roof region of the building in *Meeting Creek*. Notice that, in our method, the edge points of roof and chimneys is aligned precisely to the image.

5. ACKNOWLEDGMENT

We thank our collaborators in the Architectural Biometrics project [27] from which some datasets for our research are derived.

6. REFERENCES

[1] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

[2] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from internet photo collections," *Intl. J. Computer Vision*, vol. 80, no. 2, pp. 189–210, 2008.

[3] L. Liu and I. Stamos, "Automatic 3D to 2D registration for the photorealistic rendering of urban scenes," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, vol. 2, June 2005, pp. 137–143 vol. 2.

[4] M. Ding, K. Lyngbaek, and A. Zakhor, "Automatic registration of aerial imagery with untextured 3D lidar models," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, June 2008, pp. 1–8.

[5] T. Pylvninen, J. Berclaz, T. Korah, V. Hedau, M. Aanjaneya, and R. Grzeszczuk, "3d city modeling from street-level data for augmented reality applications," in *Intl. Conf. 3D Imaging, Modeling, Proc. Vis. Transmission*, Oct 2012, pp. 238–245.

[6] H. Cho, Y. W. Seo, B. V. K. V. Kumar, and R. R. Rajkumar, "A multi-sensor fusion system for moving object detection and tracking in urban driving environments," in *Proc. the IEEE Intl. Conf. on Robotics and Auto.*, May 2014, pp. 1836–1843.

[7] K. Ikeuchi, T. Oishi, J. Takamatsu, R. Sagawa, A. Nakazawa, R. Kurazume, K. Nishino, M. Kamakura, and Y. Okamoto, "The great buddha project: Digitally archiving, restoring, and analyzing cultural heritage objects," *Intl. J. Computer Vision*, vol. 75, no. 1, pp. 189–208, 2007.

[8] R. Pintus, E. Gobbetti, and R. Combet, "Fast and robust semi-automatic registration of photographs to 3D geometry," in *Intl. Symp. Virtual Reality, Archaeology, and Cultural Heritage*, 2011, pp. 9–16.

[9] Q. Zhang and R. Pless, "Extrinsic calibration of a camera and laser range finder (improves camera calibration)," in *IEEE Intl. Conf. Intel. Robots and Sys.*, vol. 3, Sept 2004, pp. 2301–2306 vol.3.

[10] L. Tamas and Z. Kato, "Targetless calibration of a lidar - perspective camera pair," in *IEEE Intl. Conf. Comp. Vision Workshop*, Dec 2013, pp. 668–675.

[11] G. Pandey, J. R. McBride, S. Savarese, and R. M. Eustice, "Automatic extrinsic calibration of vision and lidar by maximizing mutual information," *J. Field Robotics*, vol. 32, no. 5, pp. 696–722, 2015.

[12] A. Mastin, J. Kepner, and J. Fisher, "Automatic registration of lidar and optical images of urban scenes," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, June 2009, pp. 2639–2646.

[13] L. Liu, I. Stamos, G. Yu, G. Wolberg, and S. Zokai, "Multiview geometry for texture mapping 2D images onto 3d range data," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, vol. 2, June 2006, pp. 2293–2300.

[14] W. Zhao, D. Nister, and S. Hsu, "Alignment of continuous video onto 3D point clouds," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 27, no. 8, pp. 1305–1318, 2005.

[15] J. Castorena, U. S. Kamilov, and P. T. Boufounos, "Autocalibration of lidar and optical cameras via edge alignment," in *IEEE Intl. Conf. Acoust., Speech, and Signal Proc.*, March 2016, pp. 2862–2866.

[16] Q. Wang and S. You, "A vision-based 2D-3D registration system," in *IEEE Workshop on Appl. of Comp. Vision.*, Dec 2009, pp. 1–8.

[17] M. Bevilacqua, J. F. Aujol, M. Brdif, and A. Bugeau, "Visibility estimation and joint inpainting of lidar depth maps," in *IEEE Intl. Conf. Image Proc.*, Sept 2016, pp. 3503–3507.

[18] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, 2016.

[19] C. Wu, "Towards linear-time incremental structure from motion," in *Intl. Conf. 3D Vision*, 2013, pp. 127–134.

[20] A. Myronenko and X. Song, "Point set registration: Coherent point drift," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 32, no. 12, pp. 2262–2275, Dec 2010.

[21] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[22] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aans, "Large scale multi-view stereopsis evaluation," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, June 2014, pp. 406–413.

[23] C. Strecha, W. von Hansen, L. V. Gool, P. Fua, and U. Thoennessen, "On benchmarking camera calibration and multi-view stereo for high resolution imagery," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, June 2008, pp. 1–8.

[24] S. Katz, A. Tal, and R. Basri, "Direct visibility of point sets," in *ACM Trans. on Graphics*, vol. 26, no. 3. ACM, 2007, p. 24.

[25] G. Yang, C. V. Stewart, M. Sofka, and C.-L. Tsai, "Registration of challenging image pairs: Initialization, estimation, and decision," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 29, no. 11, pp. 1973–1989, 2007.

[26] G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama, "Digital photography with flash and no-flash image pairs," *ACM Trans. on Graphics*, vol. 23, no. 3, pp. 664–672, 2004.

[27] L. Ding, A. Elliethy, E. Freedenberg, S. A. Wolf-Johnson, J. Romphf, P. Christensen, and G. Sharma, "Comparative analysis of homologous buildings using range imaging," in *IEEE Intl. Conf. Image Proc.*, Sept 2016, pp. 4378–4382.

[28] "Architectural biometrics project website." [Online]. Available: http://www.architecturalbiometrics.com