Robust Computer Vision Techniques for High-quality 3D Modeling

Joon-Young Lee, Jiyoung Jung, Yunsu Bok, Jaesik Park, Dong-Geol Choi, Yudeog Han, and In So Kweon
Robotics and Computer Vision Lab, KAIST
{jylee, jyjung, ysbok, jspark, dgchoi, ydhan}@rcv.kaist.ac.kr, iskweon@kaist.ac.kr

Abstract—**In this paper, we present our recent sensor fusion approaches to obtain high-quality 3D information. We first discuss two fusion methods that combine geometric and photometric information. The first method, multiview photometric stereo, reconstructs the full 3D shape of a target object. The geometric and photometric information is efficiently fused by using a planar mesh representation. The second method performing shape-from-shading with a Kinect sensor estimates the shape of an object under uncalibrated natural illumination. Since the method uses a single RGB-D input, it is capable of capturing the high quality shape details of a dynamic object under varying illumination. Subsequently, we summarize a calibration algorithm of a time-of-flight (ToF) sensor and a camera fusion system with a 2.5D pattern. Lastly, we present a camera-laser sensor fusion system for the large-scale 3D reconstruction.**

I. Introduction

3D modeling is one of the most traditional problems in computer vision and graphics, and its goal is to recover 3D information of the scene. There are many approaches to obtain 3D information such as multiview stereo, structure-from-motion, shape-from-shading, and using various depth sensors. Since each approach has its own pros and cons, there is a chance for improving 3D modeling performance by fusing different approaches.

Geometric approaches such as multiview stereo and structure-from-motion give sparse metric depth, while photometric approaches such as photometric stereo and shape-from-shading give dense normal. Therefore, there have been a few successful fusion systems combining the two approaches to achieve dense and high quality 3D information. Sparse metric depth data play the role of providing absolute positions in 3D reconstruction whereas dense normals recover surface details [1], [2].

Depth sensors such as Kinect, laser scanners, and time-of-flight 3D sensors give video-rate depth information. However, Kinect and time-of-flight 3D sensors give only a rough geometry and 2D laser scanners give depth on a scan-line only. Although the depth information is not sufficient for many applications in accuracy, resolution, and field-of-view, it is shown that utilizing the rough geometric information is very useful [3].

In this paper, we present a brief overview on our recent sensor fusion techniques for high-quality 3D modeling. We present an efficient multiview photometric stereo via planar mesh representation. For practical use of high quality 3D modeling, we present a shape estimation framework from an RGB-D image under natural illumination. A calibration algorithm for a ToF-Camera fusion system based on a novel 2.5D pattern board is introduced. To reconstruct large-scale outdoor scenes, we present a camera-laser fusion system as well as a reconstruction algorithm.

II. Efficient Multiview Photometric Stereo via Planar Mesh Representation

Among various methods for recovering 3D geometry of closed-shape objects, multi-view photometric stereo (MVS) utilizes images which are taken with controlled light condition. In this section, we review our MVS algorithm [4] which is devised for recovering high-quality surface texture in an efficient manner.

A. Planar Mesh Representation for MVS

The key concept of MVS is combining photometric cues and geometric cues for detailed geometric recovery of a target object. For combining two cues in 3D domain, previous MVS methods [1], [2] utilizes 3D mesh representation which is composed of 3D vertices and surface polygons. These methods update vertices of rough geometry by using convex optimization where the penalty function evaluates inappropriate vertex positions based on photometric cues.

In this work, we propose a novel method that updates geometry in a planar mesh representation. The major benefit is threefold. 1) Multiview images can be jointly handled in a unified domain which avoids per view normal map merging. 2) The planar mesh can efficiently accommodate high-density geometry which is fit for recovering highly delicate surface details. 3) Planar representation relaxes redundant degree of freedom (DoF) for geometry update, hence the optimization become more efficient.

B. Pipeline of Proposed Method

Base Geometry Acquisition. As a first step, our method utilizes structure from motion (SfM) [5] and multi-view stereo (MVS) to acquire rough geometry and camera parameters. The detailed steps for base mesh acquisition are follows. 1) Obtain dense depth maps via semiglobal matching [6]. 2) Each depth maps are merged into voxel grid. 3) Labeling voxels using volumetric graph-cut [7]. 4) Obtain base mesh using marching cubes [8].

Mesh Parameterization. We parameterize the base mesh into 2D texture map domain for optimal fusion of photometric cues

This research was supported by the Ministry of Trade, Industry and Energy, Korea, under the Human Resources Development Program supervised by the NIPA (H1502-13-1001).

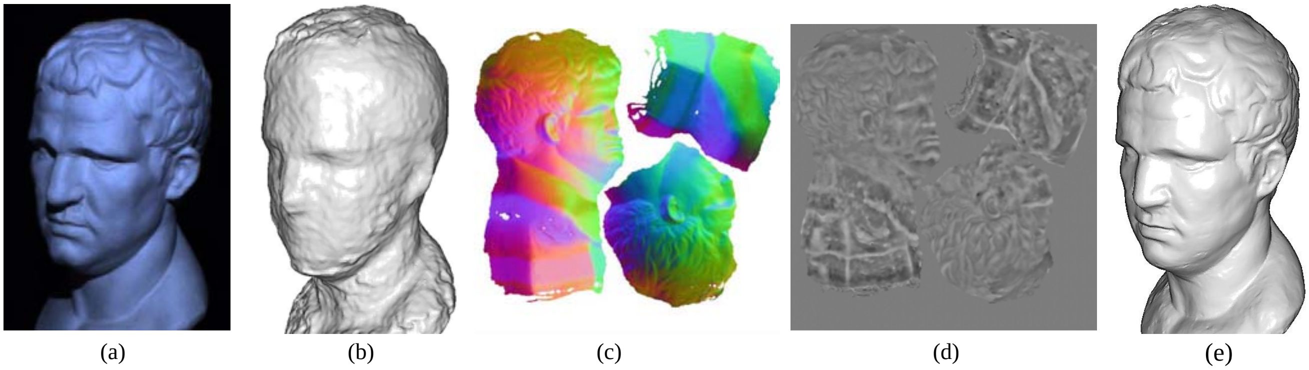


Fig. 1: Our multiview photometric stereo result. (a) One of input images. (b) Base geometry acquired by structure from motion and multiview stereo pipeline. (c) Estimated surface normal of the object in a texture domain. (d) Computed displacement map for base mesh refinement. (e) Refinement result.

and base mesh. We use Isochart [9] which minimizes non-uniform distortions which affect uniform sampling of surface mesh.

Image Warping. We warp images into texture domain with following procedure. 1) For each pixels in texture map, we find a corresponding 3D position of the base mesh. 2) Project the 3D position in image domain and get the pixel intensity. 3) Transfer the pixel intensity to the texture map.

Surface Normal Estimation. We assume dichromatic reflectance model which is combination of Lambertian shading and specular lobe. With this assumption, we give rank-3 constraint [10] on observed intensity matrix. To relieve linear ambiguity, we use surface normal from base mesh. Note that this is uncalibrated photometric stereo; the method does not require light directions for normal estimation. In addition, normal estimation is performed on the texture domain which means unified handling of multiview images.

Geometry Refinement. Given photometric normal, we refine base geometry via convex optimization. In contrast to previous methods [2], [1], we optimize geometry by solving for displacement map [11] not for 3D vertices. In this procedure the rich details in surface normal are transferred to displacement map. Our convex cost function consists of two terms. The first term measures inconsistency between normal of base mesh and estimated surface normal. The second term regularizes large displacement value. The global optimum can be computed with sparse linear matrix solver.

C. Experiment Result

Fig. 1 shows a result of our method. The object named Agrippa is placed on the dark room under controlled light condition. We use 312 images which is taken with varying viewpoints and light directions. Fig. 1 (e) shows the refined geometry by applying the displacement map to the base mesh. It shows rich level of details especially ear, hear and facial expression of the Agrippa.

III. High-quality Shape from an RGB-D Image under Natural Illumination

RGB-D sensors (e.g., Kinect) consisting of a color camera and a depth sensor become popular. While they give video-rate depth information, the depth quality is not good enough for 3D modeling applications as shown in Fig. 3. In [12], we propose a shape estimation framework that dramatically improves shape details of diffuse objects with uniform albedo from an RGB-D sensor. This section is a brief summary of the algorithm [12].

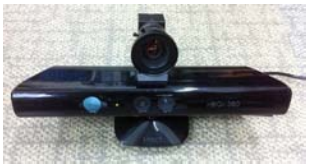


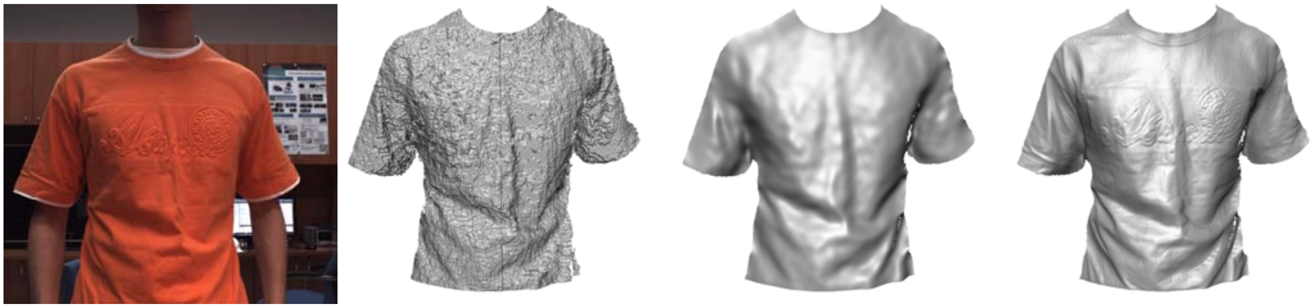
Fig. 2: System setup for real-world shape capture. Flea3 for color images + Kinect for depth images

Depth data from RGB-D sensors are typically very noisy due to the limited resolution of the depth sensor. To reduce depth noise and obtain smooth surface, we first apply the bilateral filtering on the given depth map. In the following explanation of this section, we consider depth as a smoothed one.

Our method consists of the following steps. We exploit the given color and depth to estimate a quadratic lighting model. It is followed by per-pixel lighting variable estimation that models spatially varying illumination. We determine surface normals with the estimated lighting, and then high quality shape is obtained by fusing the given geometry with the estimated normals.

Image intensity is determined by a shading function applied to the surface normal. As studied in [13], the intensity of convex diffuse objects is insensitive to high frequencies in lighting environment. Therefore, the intensity of diffuse objects can be explained by a low-dimensional global lighting model like spherical harmonics and quadratic function [14]. We use the quadratic function as a global lighting model and estimate the quadratic lighting variables from the observed intensities and the normals from the given rough geometry. Although the normals are inaccurate and possibly contain outliers, the object provides enough information for estimating the low-dimensional lighting model. Therefore we estimate the global lighting variables for each color channel by solving an over-determined linear system.

While the quadratic lighting model explains diffuse surface under natural illumination with a small number of variables,



(a) Input image (b) Input depth (c) Bilateral Smoothing (d) Our result

Fig. 3: Our result from a single RGB-D input under uncontrolled natural illumination.

there are local lighting variations due to attached shadows, interreflections and near lighting. To account for spatially varying illumination, we extend the quadratic model with multiplicative per-pixel variables.

We estimate the per-pixel lighting variables from the global estimation error, which can be decomposed into two factors; local lighting variations from the global estimate, and geometric normal deviations from the true normals. The goal of the per-pixel lighting variable estimation is to separate the local lighting variation from the total error. The geometric normal error corresponding to the missing detailed shape in the rough data is recovered in a subsequent normal optimization step.

The key idea for estimating the per-pixel variables is to exploit different frequency characteristics in the two error factors. The geometric normal deviations from the true normals have a high-frequency characteristic when compared to the accurate low-frequency structures of input depth. However, the local lighting variations from the global estimate have a low-frequency characteristic because lighting is smoothly varying [15].

With the modified quadratic lighting model, normal estimation becomes a nonlinear optimization problem. While the previous method [14] shows the nonlinear problem can be solved, it still suffers from local ambiguity, which means that the resulting surface normal is not unique. In our approach, the additional depth information greatly reduces the local ambiguity.

After normal estimation, high quality depth is obtained by fusing depth information with the estimated normals. We use the fusion algorithm described in [1] [1].

To show the performance of our method, we capture real-world objects using the Kinect-based system as shown in Fig. 2. The result of shape estimation under uncontrolled natural illumination is presented in Fig. 3. The result shows that our method successfully estimates extremely detailed shape of target objects in the presence of spatially varying illumination.

IV. Calibration of ToF and Camera Fusion System

Accurate depth estimation of the scene has been one of the key research interests for past decades. Especially in mobile robot applications, people have installed various metric depth measurement devices because accurate depth estimation is directly related to many important tasks such as mapping, navigation and obstacle avoidance. Recently, there are many kinds of depth measurement devices such as Kinect and ToF camera that can provide the full 3D of the scene in real-time. But what they provide is not in the same size as the HD color image, nor are they looking at the exact same view as the color camera. All the applications using this sensor fusion system must be followed by the exact sensor calibration. In this section, we present a review of our calibration algorithm published in [16].

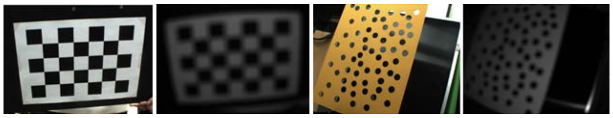


Fig. 4: A general checkerboard(first, second) and our 2.5D pattern board(third, fourth) in a color image (first, third) and an amplitude image of a ToF sensor(second, fourth)

Instead of checkerboard, we have designed a 2.5D pattern that provides correct correspondences for both color and depth images automatically. And we propose an optimization based framework for color and 3D ToF camera sensor fusion calibration. In the optimization, the error function f is to be minimized. It contains the reprojection error in the color images, and the amplitude images of the ToF camera, and the depth measurement errors of the plane in 3D. Three errors can have different weights to emphasize a certain type of error more than others in the optimization process.

A. 2.5D pattern board

One of the problems is that the commonly used checkerboard is not appropriate for the 3D ToF camera due to its low resolution. In order to get an accurate calibration result, minimizing reprojection errors using correct correspondences is crucial. But the amplitude image of a ToF camera is very blurry so that it is very hard to get the exact correspondences automatically by corner detection.

The fundamental strategy for better calibration can be simply said to find the precise correspondences between the model plane and the amplitude images of a 3D ToF camera as well as with the color images. Therefore we have constructed a 80cmx60cm pattern board that has 64 holes, as shown in Figure 3. The diameter of each hole is 4cm, which is large

[1] http://w3.impa.br/ diego/software/NehEtAl05/

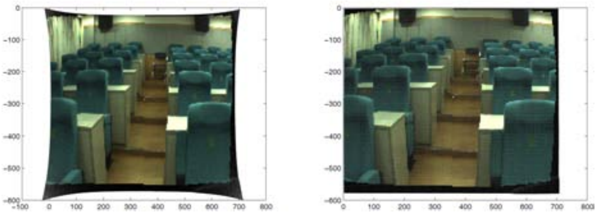


Fig. 5: Color reprojection before(left) and after(right) intrinsic parameter optimization

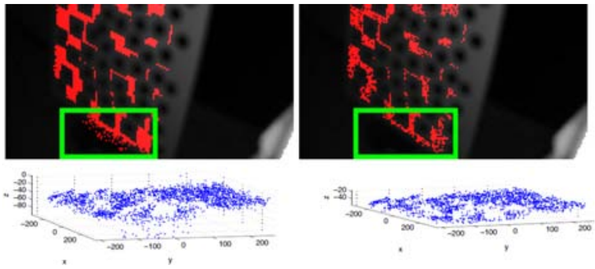


Fig. 6: Before(left) and after(right) the removal of outliers using plane fitting

enough for the infrared rays of the ToF camera to pass through the hole so that the dot patterns are clearly shown even in a 176x144-sized amplitude images.

B. Intrinsic parameter optimization

If we calculate the pinhole camera parameters using only several dot-pattern correspondences, the projection looks like the left figure in Fig. 5. The image is severely distorted especially on the boundaries because the number of correspondences is simply not enough. Therefore, we used all the 3D-2D mapping relationship gotten from a single shot, which is about 25,000 correspondences, to optimize the camera parameters including radial distortion. We used Levenberg-Marquardt optimization to minimize the reprojection errors.

C. Depth constraint

We also include depth constraint for better accuracy of the calibration result. It is the same constraint that is used for camera-laser scanner calibration. But we had to go through filtering process to outliers around the pattern plane and leave out only the depth measurements of the pattern plane to constrain them to be exact. We applied RANSAC based plane fitting. The left and right figures in Fig. 6 show before and after the removal of outliers using plane fitting, respectively.

D. Summary

In conclusion, we have presented an extrinsic calibration method to estimate the pose of a 3D ToF camera with respect to a color camera. We use 2.5D pattern so that the correct correspondences are obtained for both color and ToF cameras. For accurate reprojection error calculation, we refine the intrinsic parameters of the ToF camera to model its projection as a pinhole camera model. Depth constraint which restricts the depth measurement to lie on the pattern plane is also employed into LM optimization as well as the reprojection errors. Our process is basically fully automatic, including the acquisition of sufficient correct correspondences.

V. Large-scale 3D Reconstruction by Camera-Laser Fusion

Laser sensors provide accurate depth information. Recently 3D laser sensors such as Velodyne become popular, but 2D laser sensors are still considered as a cheap and efficient solution. In order to reconstruct 3D structures using them, they are usually mounted on a ground vehicle and scan targets vertically while the vehicle moves horizontally. If the motion of the sensor system is estimated, scanned data can be accumulated to generate a 3D point cloud of the structures. In the framework, the most important problem is to estimate the motion of the system. Usually it is assumed that the motion is defined in 2D space with 3 degrees of freedom: 1 from rotation and 2 from translation. We attach cameras to 2D laser sensors to estimate free motion in 3D space (i.e. 6 degrees of freedom). The rest of this section is a brief introduction of our sensor fusion systems published in [17] and [18].

The overall process of motion estimation consists of three steps: calibration, local estimation and global refinement. In order to utilize cameras and laser sensors in a unified algorithm, their relative poses must be computed first. Instead of using the typical point-plane constraint [19], we proposed a point-line constraint which utilizes the edges of a planar pattern. Because of the stronger constraint and accurate edge detection, calibration results by the proposed method showed higher accuracy than those by the previous method.

Local estimation step computes the relative pose between adjacent frames based on structure-from-motion (SFM) methodology. The key idea of this step is to consider scanned data as 3D points in camera coordinate system. Scanned points are transformed into camera coordinate system using the calibration result. These points are projected onto images and tracked to adjacent frame. Conventional SFM algorithms such as perspective 3-point [20] can be used to estimate the relative pose between the adjacent frames. In order to avoid degenerated configuration caused by 2D laser sensor, we proposed a new algorithm called ‘generalized laser 3-point’ [21]. In real experiments, the proposed algorithm outperformed conventional algorithms.

Frame-by-frame methods always suffer from error accumulation problem. We reduced accumulated error by closing a few loops. We registered local 3D structures to compute relative pose between two visits of the same scene. Then we optimized the poses of the loops in global coordinate system to minimize accumulated error to be distributed to frames. The accumulated error is distributed equally to all frames [22] to satisfy closed-loop constraint and maintain local accuracy simultaneously.

We designed two different systems with own purposes. The hand-held system [17] is carried by a human operator to capture and reconstruct narrow scenes. The vehicle-mounted system [18] is mounted on a ground vehicle to capture and reconstruct large-scale scenes. Reconstruction results are shown in Fig. 8.

VI. Conclusion

In this paper, we have presented a brief overview on several sensor fusion algorithms for high-quality 3D modeling. An efficient multi-view photometric stereo method can reconstruct

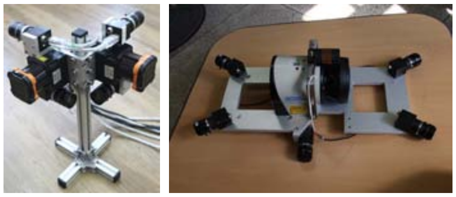
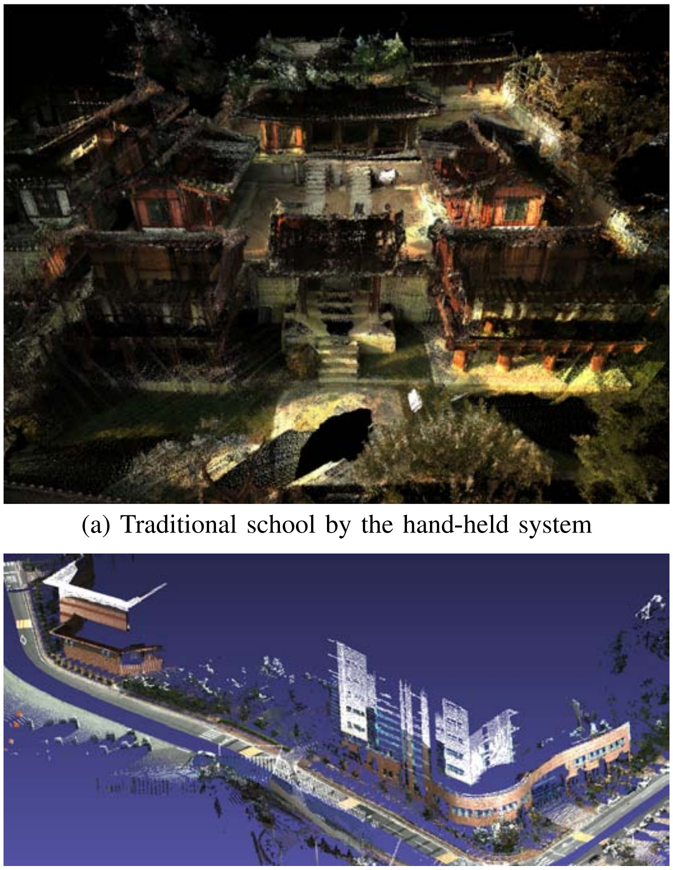


Fig. 7: Camera-laser fusion systems. (a) Hand-held system (b) Vehicle-mounted system



(a) Traditional school by the hand-held system

(b) KAIST campus by the vehicle-mounted system

Fig. 8: Reconstruction results.

the full high-quality 3D model of an object [4]. It is shown that photometric stereo combined with the well-known multiview stereo approach drastically improves the accuracy of 3D information. It is also important to note that a shape-from-shading algorithm can be a viable solution to obtain high-quality 3D information once it is given rough initial depth information by a Kinect sensor. A novel lighting model is very effective to estimate the accurate surface normal of an object under uncalibrated natural illumination [12]. Our calibration method using a novel 2.5D pattern plane makes a time-of-flight sensor and a camera fusion system practical [16]. For the reconstruction of large-scale scenes, we have presented a camera-laser sensor fusion system [17], [18]. The camera-laser fusion system have successfully reconstructed the 3D model of the KAIST campus without using GPS information. Combining photometric information with geometric information for large-scale outdoor scenes is currently underway.

References

[1] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi, "Efficiently combining positions and normals for precise 3D geometry," in *Proceedings of ACM SIGGRAPH*, 2005, pp. 536–543.

[2] C. Hernández, G. Vogiatzis, and R. Cipolla, "Multi-view photometric stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 30, no. 3, pp. 548–554, 2008.

[3] Q. Zhang, M. Ye, R. Yang, Y. Matsushita, B. Wilburn, and H. Yu, "Edge-preserving photometric stereo via depth fusion," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2472–2479.

[4] J. Park, S. Sinha, Y. Matsushita, Y.-W. Tai, and I. S. Kweon, "Multiview photometric stereo using planar mesh parameterization," in *ICCV*, 2013.

[5] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3d," in *Proceedings of ACM SIGGRAPH*, 2006, pp. 835–846.

[6] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 30, no. 2, pp. 328–341, 2008.

[7] G. Vogiatzis, C. Hernández Esteban, P. H. S. Torr, and R. Cipolla, "Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 29, no. 12, Dec. 2007.

[8] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," in *Proceedings of the 14th annual conference on Computer graphics and interactive techniques (SIGGRAPH)*, 1987, pp. 163–169.

[9] K. Zhou, J. Synder, B. Guo, and H. Shum, "Iso-charts: Stretch-driven mesh parameterization using spectral analysis," in *Eurographics Symposium on Geometry Processing*, 2004, pp. 45–54.

[10] H. Hayakawa, "Photometric stereo under a light source with arbitrary motion," *Journal of the Optical Society of America A (JOSA A)*, vol. 11, no. 11, pp. 3079–3089, 1994.

[11] L. Szirmay-Kalos and T. Umenhoffer, "Displacement mapping on the GPU - State of the Art," *Computer Graphics Forum*, vol. 27, no. 1, 2008.

[12] Y. Han, J.-Y. Lee, and I. S. Kweon, "High quality shape from a single rgb-d image under uncalibrated natural illumination," in *ICCV*, 2013.

[13] R. Ramamoorthi and P. Hanrahan, "An efficient representation for irradiance environment maps," in *Proceedings of ACM SIGGRAPH*, 2001, pp. 497–500.

[14] M. K. Johnson and E. H. Adelson, "Shape estimation in natural illumination," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 2553–2560.

[15] D. A. Forsyth, "Variable-source shading analysis," *International Journal on Computer Vision (IJCV)*, vol. 91, no. 280-302, 2011.

[16] J. Jung, Y. Jeong, J. Park, H. Ha, J. D. Kim, and I. S. Kweon, "A novel 2.5d pattern for extrinsic calibration of tof and camera fusion system," in *IROS*, 2011, pp. 3290–3296.

[17] Y. Bok, Y. Jeong, D.-G. Choi, and I. S. Kweon, "Capturing village-level heritages with a hand-held camera-laser fusion sensor," *International Journal on Computer Vision (IJCV)*, vol. 94, no. 1, pp. 36–53, 2011.

[18] Y. Bok, D.-G. Choi, Y. Jeong, and I. S. Kweon, "Capturing city-level scenes with a synchronized camera-laser fusion sensor," in *IROS*, 2011, pp. 4436–4441.

[19] Q. Zhang and R. Pless, "Extrinsic calibration of a camera and laser range finder (improves camera calibration)," in *IROS*, 2004, pp. 2301–2306.

[20] R. M. Haralick, C. N. Lee, K. Ottenberg, and M. Nolle, "Review and analysis of solutions of the three point perspective pose estimation problem," *International Journal of Computer Vision*, vol. 13, no. 3, pp. 331–356, Dec. 1994.

[21] Y. Bok, D.-G. Choi, and I. S. Kweon, "Generalized laser three-point algorithm for motion estimation of camera-laser fusion system," in *ICRA*, 2013, pp. 2865–2872.

[22] G. C. Sharp, S. W. Lee, and D. K. Wehe, "Multiview registration of 3D scenes by minimizing error between coordinate frames," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 26, no. 8, pp. 1037–1050, Aug. 2004.