A Hand-held Photometric Stereo Camera for 3-D Modeling

Tomoaki Higo[1]*, Yasuyuki Matsushita[2], Neel Joshi[3], Katsushi Ikeuchi[1]
[1]The University of Tokyo, [2]Microsoft Research Asia, [3]Microsoft Research
Tokyo, Japan, Beijing, China, Redmond, WA 98052-6399
{higo, ki}@cvl.iis.u-tokyo.ac.jp, {yasumat, neel}@microsoft.com

Abstract

This paper presents a simple yet practical 3-D modeling method for recovering surface shape and reflectance from a set of images. We attach a point light source to a hand-held camera to add a photometric constraint to the multi-view stereo problem. Using the photometric constraint, we simultaneously solve for shape, surface normal, and reflectance. Unlike prior approaches, we formulate the problem using realistic assumptions of a near light source, non-Lambertian surfaces, perspective camera model, and the presence of ambient lighting. The effectiveness of the proposed method is verified using simulated and real-world scenes.

1. Introduction

Three-dimensional (3-D) shape acquisition and reconstruction is a challenging problem with many important applications in archeology, medicine, and in the film and video game industries. Numerous systems exist for 3-D scanning using methods such as multi-view stereo, structured light, and photometric stereo; however, the use of 3-D modeling is limited by the need for large, expensive, and costly hardware setups that require extensive calibration procedures. As a result, 3-D modeling is often neither a practical nor accessible option for many applications. In this paper, we present a simple, low-cost method for object shape and reflectance acquisition using a hand-held camera with an attached point light source.

When an object is filmed with our camera setup its appearance changes both geometrically and photometrically. These changes provide clues to the shape of an object; however, their simultaneous variation prohibits the use of traditional methods for 3-D reconstruction. Standard multi-view stereo and photometric stereo assumptions fail when considered independently; however, when considered jointly their complimentary information enables high-quality shape reconstruction.

The particular concept of jointly using multi-view and photometric clues for shape acquisition is not new to this work and has become somewhat popular in recent years [23, 13, 11]; however, these previous works have several limitations that keep them from being used in practice: the need for fixed or known camera and light positions, a dark room, an orthographic camera model, and a Lambertian reflectance model. It is often difficult to fit all these constraints in real world situations, *e.g.*, to adhere to an orthographic camera and distant point light source model, one has to film the object at a distance from the camera and light, which makes hand-held acquisition impossible. Furthermore, most real-world objects are not Lambertian. Our work improves upon previous work by removing all of these constraints.

The primary contributions of this paper are: (1) an auto-calibrated, hand-held multi-view/photometric stereo camera, (2) a reconstruction algorithm that handles a perspective camera, near light configuration, ambient illumination, and specular objects, and (3) a reconstruction algorithm that performs simultaneous estimation of depth and surface normal. The rest of this paper proceeds as follows: in the next section, we will discuss the previous work in this area. In Sections 2 and 3, we discuss our algorithm. We present results in Section 4 followed by a discussion and our conclusions.

1.1. Previous work

Shape reconstruction has a long, storied history in computer vision, and, unfortunately, cannot be fully addressed within the scope of this paper. At a high-level, typical approaches use either multi-view information or photometric information separately. Multi-view stereo methods often require elaborate setups [24, 19] and, while they can excel at recovering large-scale structures, they often fail to capture high-frequency details [16]. Photometric stereo setups can be more modest, but they still require known or calibrated light positions [15] and often have inaccuracies in the low-frequencies components of the shape reconstruction [16].

Recent work has merged the benefit of these to methods using either two separate datasets [16, 22] or jointly using one dataset. Maki *et al*. [14] use a linear subspace constraint with several known correspondences to estimate light source directions up to an arbitrary invertible lin-

*This work was done while the author was visiting Microsoft Research Asia.

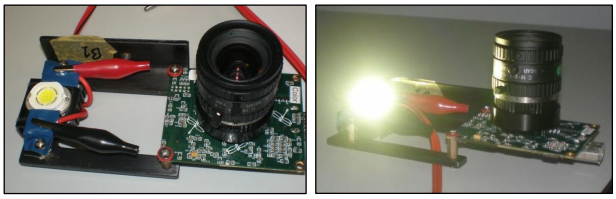


Figure 1. Our prototype implementation of the hand-held photometric stereo camera.

ear transform, but they do not recover surface normals. Simakov *et al*. [20] merge multi-view stereo and photometric constraints by assuming that the relative motion between the object and the illumination source is known. While this motion is recoverable in certain situations, there can be ambiguities. Additionally, their process can only recover normals up to an ambiguity along a plane. In contrast, our method automatically finds correspondences to recover camera parameters, with a known relative light position, and solves depth and normals without any remaining ambiguity. More recently, Birkbeck *et al*. [2] and Hernández *et al*. [10] show impressive surface reconstruction results by exploiting silhouette and shading cues using a turntable setup.

Our work is similar in spirit to that of Pollefeys *et al*. [18] who perform 3-D modeling with a perspective camera model, but use standard multi-view clues and no photometric clues, thus they do not recover normals as we do. Our work also is closely related to the work by Zhang *et al*. [23], Lim *et al*. [13], and Joshi and Kriegman [11]. Zhang *et al*. present an optical flow technique that handles illuminations changes, which requires numerous images from a dense video sequence. Lim *et al*. start with very sparse initial estimate of the shape computed from the 3-D locations for a sparse set of features and refine this shape using iterative procedure. Joshi and Kriegman extend a sparse multi-view stereo algorithm with a cost-function that uses a rank-constraint to fit the photometric variations. Our work shares some similarity with Joshi and Kriegman's approach for simultaneous estimation of depth and normals. In contrast with these three previous works, we use a known, near light position and can handle using a perspective camera and non-Lambertian objects.

2. Proposed method

Our method uses a simple configuration, *i.e*., one LED point light source attached to a camera. Fig. 1 shows a prototype of the hand-held photometric stereo camera. This configuration has two major advantages. First, it gives a photometric constraint that allows us to efficiently determine surface normals. Second, it enables a completely hand-held system that is free from heavy rigs.

Fig. 2 illustrates the flow of the proposed method. After calibrating camera intrinsics and vignetting (step 1), we take images of a scene from different view points using the camera with the LED light always turned on. Given such input images, our method first determines camera extrinsics and light source position in steps 2 and 3. In step 4, our method performs simultaneous estimation of shape, normals, albedos, and ambient lighting. We use an efficient discrete optimization to make the problem tractable. Step 5 refines the estimated surface shape by a simple optimization method. We first describe the photometric stereo formulation for our configuration in Section 2.1, and then describe the algorithmic details of our two major stages (steps 4 and 5) in Sections 2.2 and 2.3.

2.1. Near-light photometric stereo

This section formulates the photometric stereo for Lambertian objects under a near-light source with ambient illumination. Our method handles specular reflection and shadows as outliers that deviates from this formulation.

Suppose $\boldsymbol{s}$ is a light position vector that is known and fixed in the camera coordinate. Let us consider a point $\boldsymbol{x}$ on the scene surface with a surface normal $\boldsymbol{n}$ in the world coordinate. In the i-th image, the light vector $\boldsymbol{l}_i$ from the surface point $\boldsymbol{x}$ to the light source is written as

$$\boldsymbol{l}_i = \boldsymbol{s} - (\mathbf{R}_i\boldsymbol{x} + \boldsymbol{t}_i), \tag{1}$$

where $\mathbf{R}_i$ and $\boldsymbol{t}_i$ are, respectively, the rotation matrix and translation vector from the world coordinate to the camera coordinate. With the near light source assumption, intensity observation o_i is computed with accounting the inverse-square law as

$$o_i = E\rho\frac{\boldsymbol{l}_i \cdot (\mathbf{R}_i\boldsymbol{n})}{|\boldsymbol{l}_i|^3} + a, \tag{2}$$

where E is the light source intensity at a unit distance, ρ is surface albedo, and a is the magnitude of ambient illumination. Defining a scaled normal vector $\boldsymbol{b} = \rho\boldsymbol{n}$, normalized pixel intensity $o'_i = o_i/E$, and normalized ambient effect $a' = a/E$, Eq. (2) becomes

$$o'_i = \frac{\boldsymbol{l}_i \cdot (\mathbf{R}_i\boldsymbol{b})}{|\boldsymbol{l}_i|^3} + a' = \frac{(\mathbf{R}_i^T\boldsymbol{l}_i) \cdot \boldsymbol{b}}{|\boldsymbol{l}_i|^3} + a'. \tag{3}$$

Given the rotation matrix $\mathbf{R}_i$, translation vector $\boldsymbol{t}_i$, and position vector $\boldsymbol{x}$, we can easily compute the light vector $\boldsymbol{l}_i$ from Eq. (1). Once we know the light vector $\boldsymbol{l}_i$, we can estimate the scaled normal vector $\boldsymbol{b}$ on each surface point with photometric stereo. According to Eq. (3), we can compute $\boldsymbol{n}$, ρ, and a' from at least 4 observations as

$$\begin{bmatrix} o'_1 \\ o'_2 \\ o'_3 \\ o'_4 \end{bmatrix} = \begin{bmatrix} \boldsymbol{l'}_1^T & 1 \\ \boldsymbol{l'}_2^T & 1 \\ \boldsymbol{l'}_3^T & 1 \\ \boldsymbol{l'}_4^T & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{b} \\ a' \end{bmatrix}, \tag{4}$$

1. **Calibrate the Camera** (Section 3.1)
   Calibrate camera intrinsics and estimate vignetting.
2. **Estimate Camera Projection Matrices** (Section 3.2)
   Using Structure from Motion/Bundle adjustment, recover the camera projection matrices for each frame.
3. **Estimate light source position** (Section 3.2)
   Resolve the scale ambiguity by using our photo consistency on feature points from the structure from motion process.
4. **Compute Dense Depth and Normal Map** (Section 2.2)
   Find the dense depth map and normals by minimizing our near light-source, multi-view photometric constraint using a graph cut.
5. **Compute Final Surface** (Section 2.3)
   Recover the final surface by fusing the recovered dense depth map and normal field.

Figure 2. Our shape reconstruction algorithm.

where we define the near light vector $\boldsymbol{l}'_i = \mathbf{R}_i^T \boldsymbol{l}_i / |\boldsymbol{l}_i|^3$. By solving the linear system, we can estimate $\boldsymbol{n}$, ρ, and a'.

The above derivation shows how to recover normals using near-light source photometric stereo once image correspondence is known; however, for our setup where we want to leverage multi-view clues, correspondence is unknown and must be estimated. Estimating the unknown correspondence is one of the key concerns of this work and is discussed in the next section.

2.2. Simultaneous estimation of depth and normal

Our method simultaneously estimates depth, normal, surface albedo, and ambient lighting. To do this we estimate correspondence to get position information and use photometric clues to get normals – these two are fused to get the final depth. To compute correspondence, we run a stereo algorithm, where we replace the traditional match function that uses brightness constancy with one that uses the photometric clues, normal consistency, and surface smoothness. We formulate the problem in a discrete optimization framework.

Let us first assume the camera positions and light position are known – the estimation of these parameters is discussed in detail in Section 3.2. Suppose that we have m images taken from different view points with our camera. We recover correspondence by performing plane-sweep stereo. For each depth in the plane-sweep, we warp the set of images from different view points to align to one reference view. In this reference camera coordinate frame, the depth planes are assumed in the z direction parallel to the xy plane at a regular interval Δ_z.

Specifically, we warp each image to the reference camera coordinate for depth $z_j = z_0 + j\Delta_z$ using a 2-D projective transform $\boldsymbol{H}_{ij}$ as

$$\boldsymbol{p}_w = \boldsymbol{H}_{ij} \boldsymbol{p}_o, \tag{5}$$

where $\boldsymbol{p}_w$ and $\boldsymbol{p}_o$ represent the warped pixel location and the original pixel location, respectively, described by $\boldsymbol{p} = [u\, v\, 1]^T$ in the image coordinate system. Then we perform an optimization over this set of warped images to find the optimal per-pixel depth z_j that gives the best agreement among the registered pixels (given pixel p in the reference view and corresponding pixels in the warped images $I_{ij}(p) \quad (i = 1, 2, \ldots, m)$). This is done according to three criteria: photo consistency, a surface normal constraint, and a smoothness measure.

Photo consistency. Our photo consistency measure is defined to account for varying lighting, since the light source is attached to the moving camera. To explicitly handle shadows, specular reflections, and occlusions, we use a RANSAC [8] approach to obtain the initial guess of surface normal $\boldsymbol{n}_p$, surface albedo $\boldsymbol{\rho}_p$, and ambient $\boldsymbol{a}_p$ using the near-light photometric stereo assumption described in Section 2.1. The vector form of surface albedo $\boldsymbol{\rho}_p$ and ambient $\boldsymbol{a}_p$ contain elements of three color channels. Using the initial guess, the photo consistency g is checked with each of other $m - 4$ images at a given pixel p as

$$g_i(\boldsymbol{n}_p, \boldsymbol{\rho}_p, \boldsymbol{a}_p) = \sum_{c=\{R,G,B\}} |I_i^c(p) - E^c \rho_p^c \boldsymbol{l'} \cdot \boldsymbol{n}_p - a_p^c|. \tag{6}$$

We also compute the number of images that satisfy the photo consistency N as

$$N = |\{i \,|\, g_i(\boldsymbol{n}_p, \boldsymbol{\rho}_p, \boldsymbol{a}_p) < \tau\}|, \tag{7}$$

where τ is a threshold for photo consistency. The RANSAC process above computation is repeated to find the best estimates of $\boldsymbol{n}_p$, $\boldsymbol{\rho}_p$, and $\boldsymbol{a}_p$ that maximizes N at each p and depth label j. Finally, the photo consistency cost E_p is evaluated as

$$E_p(p, j) = \eta \frac{1}{N} \sum_{i \in N} g_i(\boldsymbol{n}_p, \boldsymbol{\rho}_p, \boldsymbol{a}_p) - N, \tag{8}$$

where η is a scaling constant. The first term in the cost function assesses the overall photo consistency, and the second term evaluates the reliability of the photo consistency, *i.e.*, when it is supported by many views (number of N), it is more reliable. These two criteria are combined together using a scaling constant term η. In our implementation, we fixed η as $\eta = 1/\tau$.

Surface normal constraint. Preferred depth estimates are those which are consistent with the surface normal estimates. We use a surface normal cost function $E_n(p,j)$ to enforce this criterion. Let j' be the depth label of the neighboring pixel p' that is located nearest in 3-D coordinates to the plane specified by the site (p,j) and its surface normal. Sometimes, the site (p',j') does not have a valid surface normal due to unsuccessful fitting of a surface normal by RANSAC. In that case, we take the next nearest site as (p',j'). Once the appropriate j' is found within $|j-j'|<T_j$, a vector $\boldsymbol{d}_{(p,j)}^{(p',j')}$ that connects (p,j) and (p',j') in the 3-D coordinate is defined on the assumed plane. We then compute the agreement of the surface normal at (p',j') with the depth estimate by evaluating if these two vectors are perpendicular to each other. The surface normal cost function is defined as

$$E_n(p,j)=\begin{cases}\sum_{p'}(|j-j'|+1)\boldsymbol{n}_{p'j'}\cdot\boldsymbol{d}_{(p,j)}^{(p',j')} & \text{if } |j-j'|<T_j\\ C_0\ (=\text{const.}) & \text{otherwise.}\end{cases}, \tag{9}$$

Smoothness constraint. We use a smoothness constraint on depth to penalize large discontinuities. Suppose p and p' are neighboring pixels whose depth labels are j and j' respectively. The smoothness cost function E_s is defined as

$$E_s(j,j')=|z_j-z_{j'}|=\Delta_z|j-j'|. \tag{10}$$

Energy function. Finally, the energy function E is defined by combining above three constraints as

$$E(p,j,j')=E_p(p,j)+\lambda_n E_n(p,j)+\lambda_s E_s(j,j'). \tag{11}$$

We use a 2-D grid graph cut framework to optimize the energy function. The 2-D grid corresponds to the pixel grid, *i.e.*, we define each pixel p as a site and the depth label j is associated. We use Boykov *et al.* [5, 12, 4]'s graph cut implementation to solve the problem. By solving Eq. (11), we obtain the estimates of depth, surface normal, surface albedo, and ambient lighting.

2.3. Refinement of surface shape

The depth estimate obtained by the solution method described in the previous section is discretized, and therefore it is not completely accurate due to the quantization error. To refine the depth estimate, we perform a regularized minimization of a position error, normal constraint, and smoothness penalty, to derive the optimal surface Z. The optimization method is based on Nehab *et al.* [16], and we define the error function following the work of Joshi and Kriegman [11]:

$$J(Z)=E^P+E^N+E^S. \tag{12}$$

The position error E^P is the sum of squared distances between the optimized positions S_p and original positions S'_p in the 3-D coordinate:

$$E^P=\lambda_1\sum_p||S_p-S'_p||^2, \tag{13}$$

where λ_1 is the relative weighting of the position constraint versus the normal constraint. To evaluate the position error, depth values are transformed to distances from the center of the perspective projection:

$$\begin{aligned}||S_p-S'_p||^2 &= \mu_p^2(z_p-z'_p)^2, \\ \mu_p^2 &= \left(\frac{x}{f_x}\right)^2+\left(\frac{y}{f_y}\right)^2+1,\end{aligned} \tag{14}$$

where f_x and f_y are the camera focal lengths in pixels, and z'_p is the depth value of the original position p'.

The normal error constrains the tangents of the final surface to be perpendicular to the input normals:

$$E^N=(1-\lambda_1)\sum_p\Big(\big(\boldsymbol{n}_p\cdot\boldsymbol{T}_p^x\big)^2+\big(\boldsymbol{n}_p\cdot\boldsymbol{T}_p^y\big)^2\Big), \tag{15}$$

where $\boldsymbol{T}_p^x$ and $\boldsymbol{T}_p^y$ represent the tangent vectors:

$$\begin{aligned}\boldsymbol{T}_p^x &= \Big[-\frac{1}{f_x}\Big(x\frac{\partial Z_p}{\partial x}+Z_p\Big),-\frac{1}{f_y}y\frac{\partial Z_p}{\partial x},\frac{\partial Z_p}{\partial x}\Big]^T,\\ \boldsymbol{T}_p^y &= \Big[-\frac{1}{f_x}x\frac{\partial Z_p}{\partial y},-\frac{1}{f_y}\Big(y\frac{\partial Z_p}{\partial y}+Z_p\Big),\frac{\partial Z_p}{\partial y}\Big]^T.\end{aligned}$$

The smoothness constraint penalizes high second-derivatives by penalizing the Laplacian of the surface:

$$E^S=\lambda_2\sum_p\nabla^2 Z_p. \tag{16}$$

λ_2 is a regularization parameter to control the amount of smoothing.

Each pixel generates at most 4 equations: one for the position error, one for the normal error in each of x and y directions, and one for the smoothness. Therefore, the minimization can be formulated as a large, sparse over-constrained system to be solved by least squares:

$$\begin{bmatrix}\lambda_1\mathcal{I}\\ (1-\lambda_1)\mathcal{N}\cdot\mathcal{T}^x\\ (1-\lambda_1)\mathcal{N}\cdot\mathcal{T}^y\\ \lambda_2\nabla^2\end{bmatrix}[Z]=\begin{bmatrix}\lambda_1 z\\ 0\\ 0\\ 0\end{bmatrix}, \tag{17}$$

where $\mathcal{I}$ is an identity matrix and $\mathcal{N}\cdot\mathcal{T}^x$ and $\mathcal{N}\cdot\mathcal{T}^y$ are matrices that, when multiplied by the unknown vector Z, evaluate the normal constraints $(1-\lambda_1)\boldsymbol{n}\cdot T^x$ and $(1-\lambda_1)\boldsymbol{n}\cdot T^y$. We solve this system using a conjugate gradient method for sparse linear least squares problems [17].

3. Implementation

3.1. Calibration

Before data acquisition, we calibrate the intrinsic parameters of the camera and vignetting. We use Camera Calibration Toolbox for Matlab [3] to estimate the camera intrinsics. For vignetting correction, we take images under a uniform illumination environment with a diffuser to create a vignetting mask. During the data acquisition, we move the camera system with the LED light on, without changing the intrinsic parameters of the camera.

3.2. Structure from motion

From the image sequence, we use the state-of-the-art structure from motion implementation *Bundler* [21] to estimate camera extrinsics and 3-D positions of feature points.

Unfortunately, the estimated 3-D positions of feature points have a scaling ambiguity because of the fundamental ambiguity of structure from motion. The scale k can affect the light vector estimation in Eq. (1) as

$$\boldsymbol{l}_i = \boldsymbol{s} - k(\mathbf{R}_i\boldsymbol{x} + \boldsymbol{t}_i). \tag{18}$$

We resolve this ambiguity using our photo consistency measure on feature points $\mathcal{F}$. The photo consistency cost E_p of Eq. (8) varies with the scaling parameter k. We find the optimal k that minimizes the score of $E_p(k)$ using the feature points $\mathcal{F}$ as

$$E_p(k) = \sum_{p\in\mathcal{F}} \Big[\eta \frac{1}{N} \sum_i g_i(\boldsymbol{n}_p, \boldsymbol{\rho}_p, \boldsymbol{a}_p) - N\Big]. \tag{19}$$

We minimize $E_p(k)$ by simply sweeping the parameter space of k to obtain the solution.

3.3. Coarse-to-fine implementation

The simultaneous estimation method described in Section 2.2 gives good estimates; however, the computational cost becomes high when the image resolution is large and also when many depth labels are considered. We adopt a coarse-to-fine approach to avoid this issue.

First, image pyramids are created for the registered images after image warping by Eq. (5). At the coarsest level, the simultaneous estimation method is applied using full depth labels. In the finer level of the pyramid, we expand the depth labels from the earlier level and use them as the initial guess. From this level, we prepare only a small range of depth labels around the initial guess for each site p. Using the minimum and maximum depth labels, $j_{\min}$ and $j_{\max}$, of the site and its neighboring sites, the new range is defined as $[j_{\min} - 1, j_{\max} + 1]$. We also use a finer Δ_z in the finer level of the pyramid. We set $\Delta_z \leftarrow \Delta_z/2$ when moving to the finer level of the pyramid.

	Depth [%]		Normal [deg.]		Albedo	
	mean	med	mean	med	mean	med
Baseline	**1.73**	**0.42**	10.5	**4.27**	**0.05**	**0.02**
Textureless	3.05	0.46	11.2	4.74	**0.05**	**0.02**
Specular	1.77	**0.42**	**10.0**	4.63	**0.05**	0.03
Ambient	2.68	0.47	**10.0**	4.44	**0.05**	**0.02**

Table 1. Quantitative evaluation using synthetic scenes. "mean" and "med" indicate mean and median errors, respectively.

4. Experiments

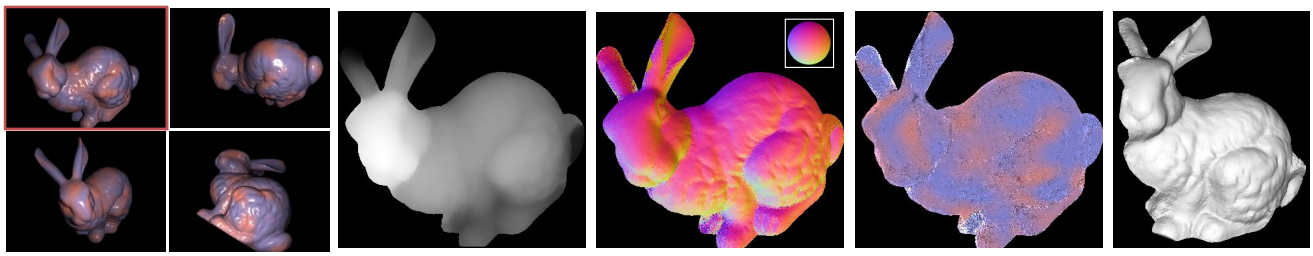
We use a Point Grey DragonFly camera (640×480) with an attached point light source as our prototype system. The camera can sequentially capture images, and we use this capability for the ease of data acquisition. During the capturing, the point light source is always turned on.

In this section, we first show quantitative evaluation using synthetic data in Section 4.1. We use three real-world scenes that have different properties to verify the applicability of the proposed method in Section 4.2. We further show comparisons with other state-of-the-art 3-D modeling methods using the real-world scenes. Throughout the experiments, we use $\tau = [6.0, 8.0]$, $\lambda_n = 7.5$ and $\lambda_s = [1.5, 3.0]$, $\lambda_1 = [0.01, 0.1]$, $\lambda_2 = [0.5, 1.5]$, $C_0 = 5$, and initial $\Delta_z = 8.0[\mathrm{mm}]$.

4.1. Simulation results

In the simulation experiments, we render synthetic scenes by simulating the configuration of our photometric stereo camera. We created a baseline scene which is textured, Lambertian, and has no ambient lighting. By changing the settings so that the objects were (1) textured, (2) have specular reflectance, and (3) the scene has ambient lighting, we assess the performance variation in comparison with the baseline case.

Table 1 shows the summary of the evaluation. From top to bottom, the results of the baseline, textureless, specular, and ambient cases are shown. The errors are evaluated using the ground truth depth map, normal map, and albedo map by looking at the mean and median errors. The depth error is represented by percentage, using [maximum depth - minimum depth] as 100%. The surface normal error is evaluated by the angular error in degrees, and albedo error is computed by taking the average of the absolute difference in R, G, and B channels, in the normalized value range $[0, 1]$. The mean error is sensitive to outliers, while the median error is not. Looking at the median error, the estimation accuracy is quite stable across the table. The textureless case produces slightly larger errors, and this indicates that there still remains ambiguous matchings even with the near light source configuration. Fig. 3 shows the result on the simulated scene with specularity.



Input images Depth map Normal map Albedo Rendering

Figure 3. Simulation result using the bunny scene. From left to right, input images (reference view in the top-left), the estimated depth map, normal map, albedo, and a final rendering of the surface are shown. In the depth map, brighter is nearer and darker is further from the camera. In the normal map, a reference sphere is placed for better visualization. 62 images are used as input.

4.2. Real-world results

We applied our method to various different real-world scenes. We show three scenes: (1) statue scene (textureless, roughly Lambertian), (2) bag scene (textured, glossy surfaces), and (3) toy scene (various reflectance properties, complex geometry).

Fig. 4 shows the result of statue scene. To produce the result, we manually masked out the background portion of the statue in the reference image. Our method can recover the surface and normal map as well as surface albedo from a textureless scene. Fig. 7 and Fig. 8 show the results of the bag scene and toy scene, respectively. These scenes contain textured surfaces as well as specularities. Our method can handle these cases as well because of our robust estimation scheme to handle specularities. Our handheld camera is particularly useful for measuring scenes like the toy scene that are difficult to move to a controlled setup.

To demonstrate the effectiveness of our photometric constraint, we have performed a comparison with a state-of-the-art multi-view stereo method proposed by Goesele *et al*. [9] that does not use a photometric constraint. The input data is obtained by fixing a camera at each view point and capturing two images with the attached point light source on and off. The images without the point light source but under environment lighting are used as input for Goesele *et al*.'s method. Fig. 5 shows the rendering of two surfaces recovered by our method and Goesele *et al*.'s method. Typical multi-view stereo algorithms can only establish a match in areas with some features (texture, geometric structure, or shadows), and this example is particularly difficult for them as it lacks such features in the most of the areas. On the other hand, our method works well because of the photometric constraint.

We also compare our method to a result from Joshi and Kriegman's method [11]. In their method, far-distant lighting and orthographic projection are assumed. We use the same dataset from their experiment and approximate their assumptions by diminishing light-fall off term $(1/|\mathbf{l}_i|^2)$ in Eq. (2) and using large focal lengths f_x and f_y. The side-by-side comparison is shown in Fig. 6. Our method can produce a result with equal quality to their method.

Our method Goesele *et al*.'s method [9]

Figure 5. Comparison with a multi-view stereo method without a photometric constraint [9] using the statue scene. 93 images are used as input for both methods.

Input image JK [11] Our method

Figure 6. Comparison with Joshi and Kriegman's method (JK) using the cat scene. Eight images are used as input for both methods. Note that rendering parameters are different as the original parameters are not available.

5. Discussion and Future Work

We presented a simple, low-cost method for high-quality object shape and reflectance acquisition using a hand-held camera with an attached point light source. Our system is more practical than those in previous work and can handle

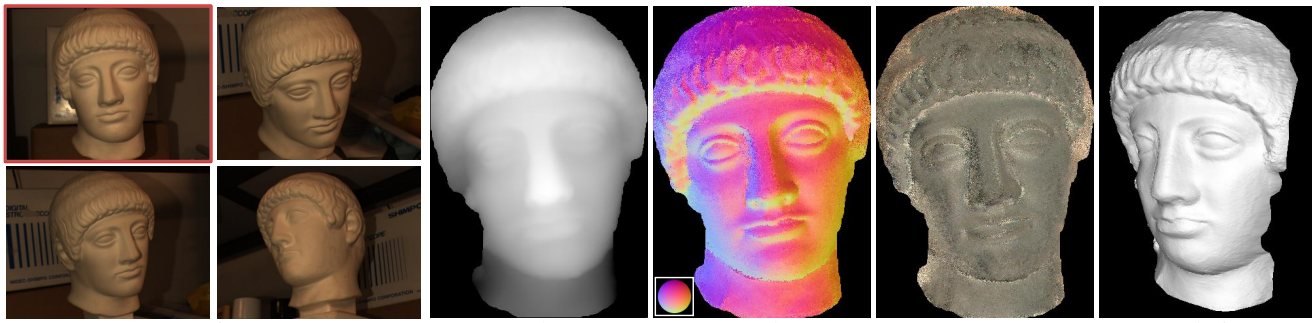


Figure 4. Result of the statue scene. From left to right, input images (reference view in the top-left), the estimated depth map, normal map, albedo, and a final rendering of the surface are shown. 93 images are used as input.

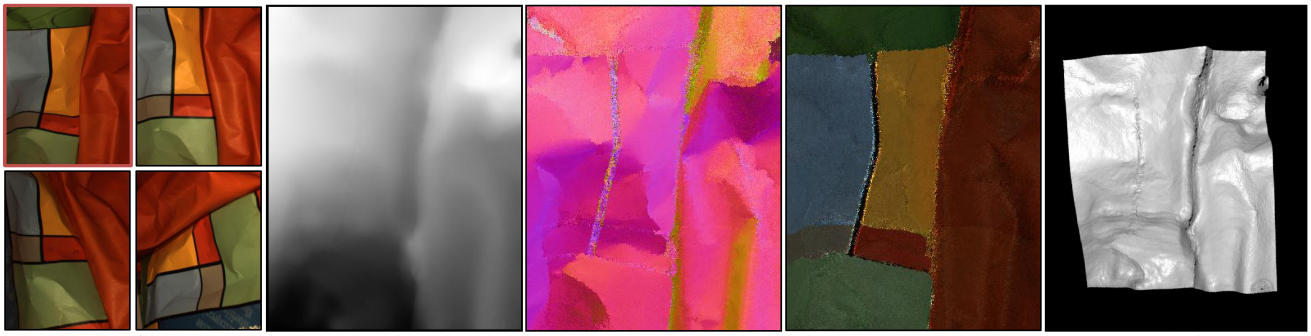


Figure 7. Result of the bag scene. From left to right, input images (reference view in the top-left), the estimated depth map, normal map, albedo, and a final rendering of the surface are shown. 65 images are used as input.

hand-held filming scenarios with a broad range of objects under realistic filming conditions. Nevertheless, there are some limitations and several avenues for future work.

One current limitation is that we only implicitly account for self-occlusions, shadowing, inter-reflections, and specularities. Our robust fitting method addresses these properties by treating them all as outliers from a Lambertian shading model. While this works well in practice, it is very likely that explicitly accounting for these factors would improve our results. We are investigating methods that could be used to explicitly model outlier pixels as self-occlusions, shadows, and inter-reflections [1, 7, 6] and methods to fit an appearance model to specularities in the data. Not only would this help refine the 3-D shape and reflectance model, it should enable higher quality rendering of scanned objects.

Another direction for future work is to perform a full 3-D reconstruction. Currently, we produce a single hight-field for a selected reference view. We are very interested using either a two-stage process of producing and merging multiple height maps into a 3-D model [9] or performing our optimization directly in the 3-D space.

References

[1] S. Barsky and M. Petrou. The 4-source photometric stereo technique for three-dimensional surfaces in the presence of highlights and shadows. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(10):1239–1252, 2003.

[2] N. Birkbeck, D. Cobzas, P. Sturm, and M. Jagersand. Variational Shape and Reflectance Estimation under Changing Light and Viewpoints. *Proc. of European Conf. on Computer Vision*, 2006.

[3] J. Y. Bouguet. Camera calibration toolbox for matlab. Technical report, 2007. Software available at http://www.vision.caltech.edu/bouguetj/calib_doc/.

[4] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004.

[5] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[6] M. Chandraker, S. Agarwal, and D. Kriegman. ShadowCuts: Photometric Stereo with Shadows. *Proc. of Computer Vision and Pattern Recognition*, 2007.

[7] M. K. Chandraker, F. Kahl, and D. J. Kriegman. Reflections on the generalized bas-relief ambiguity. In *Proc. of Computer Vision and Pattern Recognition*, pages 788–795, 2005.

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

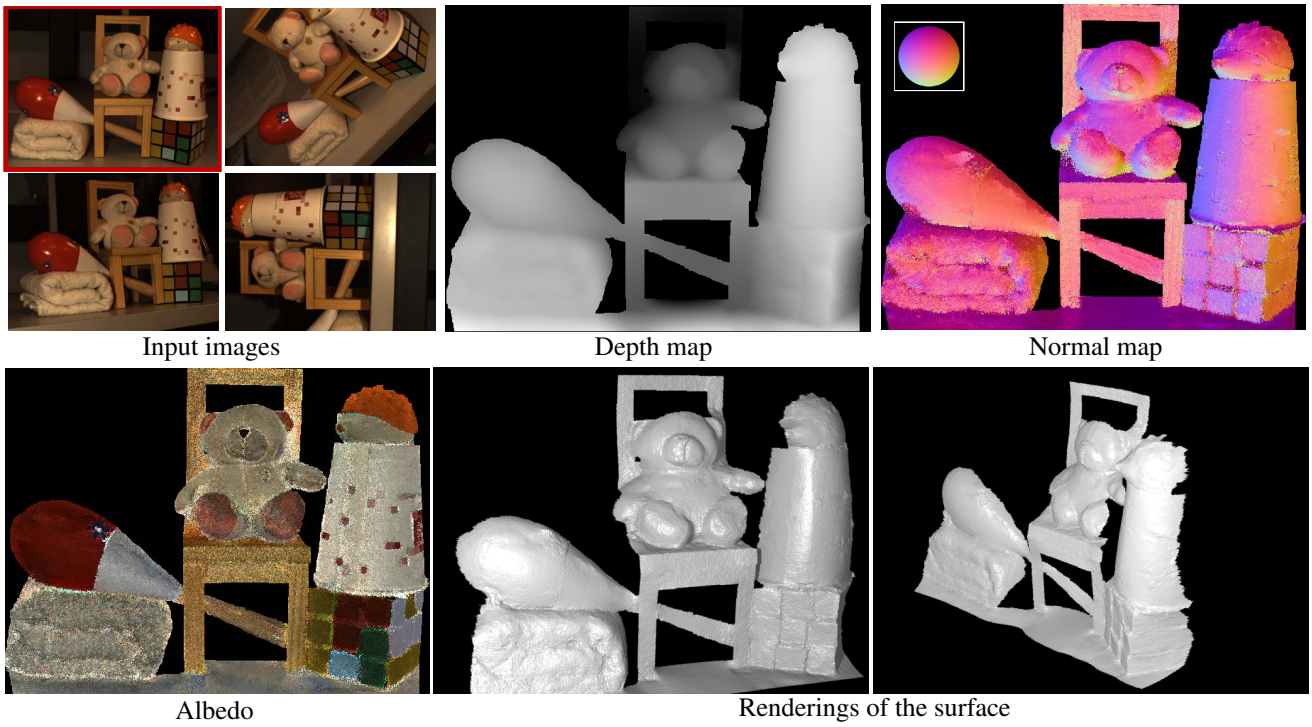


Figure 8. Result of the toy scene. The scene contains various color and reflectance properties. On the top row, from left to right, a few input images (reference view in the top-left), estimated depth map, and normal map are shown. The bottom row shows the estimated albedo map and renderings of the final surface. 84 images are used as input.

[9] M. Goesele, B. Curless, and S. Seitz. Multi-view stereo revisited. *Proc. of Computer Vision and Pattern Recognition*, 2:2402–2409, 2006.

[10] C. Hernández, G. Vogiatzis, and R. Cipolla. Multiview photometric stereo. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 3(30):548–554, 2008.

[11] N. Joshi and D. Kriegman. Shape from Varying Illumination and Viewpoint. *Proc. of Int'l Conf. on Computer Vision*, pages 1–7, 2007.

[12] V. Kolmogorov and R. Zabih. What Energy Functions Can Be Minimized via Graph Cuts? *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 147–159, 2004.

[13] J. Lim, J. Ho, M. Yang, and D. Kriegman. Passive Photometric Stereo from Motion. *Proc. of Int'l Conf. on Computer Vision*, 2:1635–1642, 2005.

[14] A. Maki, M. Watanabe, and C. Wiles. Geotensity: Combining Motion and Lighting for 3D Surface Reconstruction. *Int'l Journal of Computer Vision*, 48(2):75–90, 2002.

[15] T. Malzbender, B. Wilburn, D. Gelb, and B. Ambrisco. Surface enhancement using real-time photometric stereo and reflectance transformation. *Proceedings of EGSR*, 2006.

[16] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3D geometry. *Proc. SIGGRAPH*, 24(3):536–543, 2005.

[17] C. Paige and M. Saunders. LSQR: An Algorithm for Sparse Linear Equations and Sparse Least Squares. *ACM Trans. on Mathematical Software (TOMS)*, 8(1):43–71, 1982.

[18] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch. Visual Modeling with a Hand-Held Camera. *Int'l Journal of Computer Vision*, 59(3):207–232, 2004.

[19] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. *Proc. of Computer Vision and Pattern Recognition*, 1:519–526, June 2006.

[20] D. Simakov, D. Frolova, and R. Basri. Dense shape reconstruction of a moving object under arbitrary, unknown lighting. In *Proc. of Int'l Conf. on Computer Vision*, pages 1202–1209, 2003.

[21] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3d. *Proc. SIGGRAPH*, pages 835–846, 2006. http://phototour.cs.washington.edu/bundler/.

[22] T. Weyrich, W. Matusik, H. Pfister, B. Bickel, C. Donner, C. Tu, J. McAndless, J. Lee, A. Ngan, H. W. Jensen, and M. Gross. Analysis of human faces using a measurement-based skin reflectance model. *ACM Trans. Graph.*, 25(3):1013–1024, 2006.

[23] L. Zhang, B. Curless, A. Hertzmann, and S. Seitz. Shape and motion under varying illumination: unifying structure from motion, photometric stereo, and multiview stereo. *Proc. of Int'l Conf. on Computer Vision*, pages 618–625, 2003.

[24] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. *ACM Trans. Graph.*, 23(3):600–608, 2004.