Int J Comput Vis
DOI 10.1007/s11263-012-0589-5

A 3D Imaging Framework Based on High-Resolution Photometric-Stereo and Low-Resolution Depth

Zheng Lu · Yu-Wing Tai · Fanbo Deng · Moshe Ben-Ezra · Michael S. Brown

Received: 5 November 2011 / Accepted: 10 October 2012

Abstract This paper introduces a 3D imaging framework that combines high-resolution photometric stereo and low-resolution depth. Our approach targets imaging scenarios based on either macro-lens photography combined with focal stacking or a large-format camera that are able to image objects with more than 600 samples per mm^2. These imaging techniques allow photometric stereo algorithms to obtain surface normals at resolutions that far surpass corresponding depth values obtained with traditional approaches such as structured-light, passive stereo, or depth-from-focus. Our work offers two contributions for 3D imaging based on these scenarios. The first is a multi-resolution, patched-based surface reconstruction scheme that can robustly handle the significant resolution difference between our surface normals and depth samples. The second is a method to improve the initial normal estimation by using all the available focal information for images obtained using a focal stacking technique.

Keywords 3D Reconstruction · High resolution · Photometric stereo · Focal stack

1 Introduction

Impressive gains by digital imaging sensors are allowing photometric stereo techniques to estimate surface normals with resolutions that far surpass that possible with conventional 3D imaging techniques, e.g. structured-light, time-of-flight scanners, stereo vision. To help appreciate this difference, Fig. 1 shows an example of high-resolution surface normals compared to the 3D geometry captured using a standard structured-light scanner and a high-end commercial laser scanner designed for industrial inspection (Fig. 1a–c). The estimated surface normals exhibit significantly more details of the object's surface than that obtained by structured-light and the laser-scanner.

Photometric stereo, however, only provides 2.5D information in the form of surface normals. A common procedure is to leverage surface normal details with 3D depth values by combining the data (e.g. Bernardini et al. 2002; Nehab et al. 2005; Vlasic et al. 2009). Such prior techniques however have not been able to accommodate such vast differences in the sampling rate between the estimated surface normals and 3D geometry. Our goal is to be able to address resolution differences in the order of 100:1. For example, Fig. 1d shows our results obtained by combining the normals from Fig. 1a sampled at over 600 samples per mm^2 with the structured-light scanner sampled at roughly 6 samples per mm^2. To our best knowledge, this represents 3D data with some of the highest sampling rate demonstrated to date.

Z. Lu · F. Deng · M. S. Brown
National University of Singapore, Computing 1,
13 Computing Drive, Singapore 127110, Singapore

F. Deng
e-mail: dfanbo@comp.nus.edu.sg

M. S. Brown
e-mail: brown@comp.nus.edu.sg

Y.-W. Tai
Department of Computer Science, Korea Advanced Institute of Science and Technology, Guseong-dong, Yuseong-gu,
Daejeon 305-701, South Korea
e-mail: yuwing@cs.kaist.ac.kr

M. Ben-Ezra
Microsoft Research Asia, Building 2, No. 5 Dan Ling Street,
Haidian District, Beijing 10080, People's Republic of China
e-mail: mosheb@microsoft.com

Present Address:
Z. Lu (✉)
Department of Computer Science, University of Texas at Austin, Austin,
TX 78712, USA
e-mail: luzheng@cs.utexas.edu

Published online: 09 November 2012

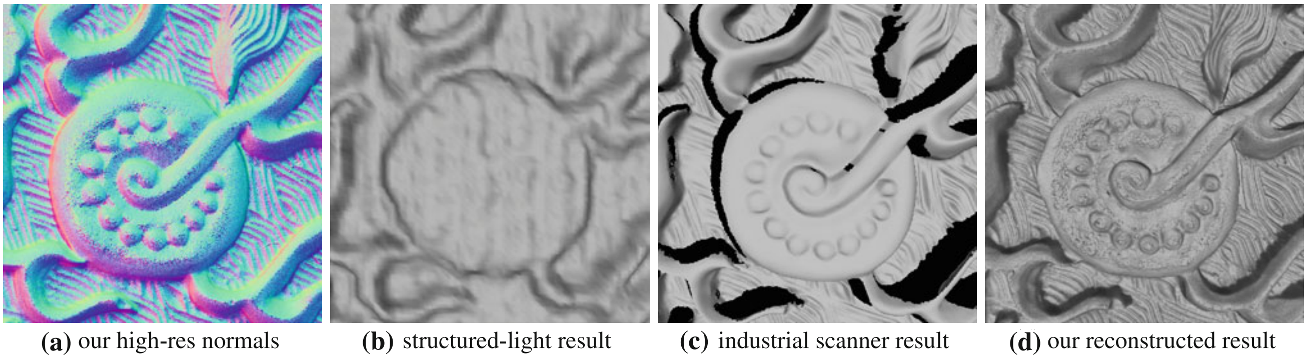
(a) our high-res normals
(b) structured-light result
(c) industrial scanner result
(d) our reconstructed result

Fig. 1 Comparison of our results against a standard structure-light scanner and a state-of-the-art industrial 3D scanner: **a** our input from photometric stereo (over 600 samples per mm^2), **b** surface reconstructed from a standard structured-light system (6.25 samples per mm^2), **c** surface reconstructed by a Konica Minolta Range 7 (168 samples per mm^2), **d** our reconstructed high-res surface

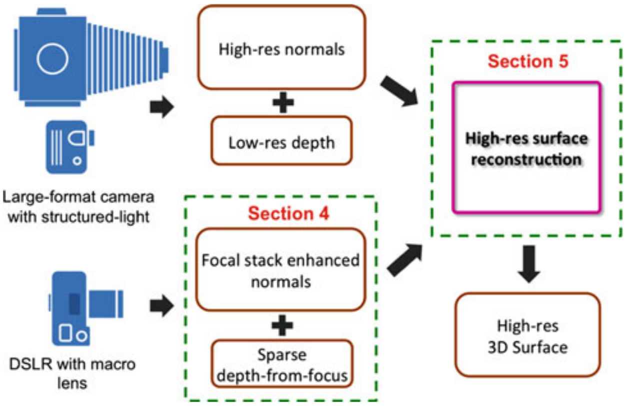

Fig. 2 Overview of our framework. Both systems, i.e. the large-format camera with structured-light and the DSLR with macro lens, produce high-resolution surface normals and sparse/low-resolution depth. Our surface reconstruction algorithm fuses these inputs together. Our normal regularization using focal stacking and depth-from-focus estimation are described in Sect. 4. Our surface reconstruction algorithm is presented in Sect. 5

In this paper, we examine two ways to obtain such high-resolution data as shown in Fig. 2. The first uses a large-format digital camera, while the second relies on a conventional digital single-lens reflex camera (DSLR) with a macro lens. In the latter case, the object is required to be much closer to the camera in order to obtain high-resolution data, leading to a shallow depth-of-field. As a result, capturing multiple images at varying focal lengths, i.e. *focal stacking*, is required to extend the depth-of-field to capture the object. This scenario has interesting implications as we now have significantly more input images with a conventional setup.

Our main contribution of this work is to propose a multi-resolution surface reconstruction scheme that fuses the low-resolution geometric data with the photometric stereo data at increasing levels of details. To deal with the large amount of data from the high-resolution input, we adopt a patch-based scheme that uses an additional boundary constraint to maintain patch coherence at the boundaries. The results of our approach are 3D surfaces captured at an exceptionally high level of detail. A secondary contribution is to demonstrate how to improve the normal estimation by globally optimizing the estimated normals against focal stack images. In addition, we show how we can use photometric lighting to improve depth-from-focus results that can be used as geometric data in our surface reconstruction algorithm. While our normal refinement algorithm assumes a Lambertian surface, our multi-resolution surface reconstruction approach directly processes surface normals and geometry and can therefore be applied with any photometric stereo approaches.

A shorter version of this work appeared in Lu et al. (2010) which only focused on the imaging scenario with a large-format digital camera. This journal version extends our conference work with novel normal refinement algorithm using focal stack in Sect. 4. Additional experimental results are shown in Sect. 6.

The remainder of this paper is organized as follows: Sect. 2 discusses related works; Sect. 3 describes our system setup; Sect. 4 describes our normal regularization using focal stacking and depth-from-focus estimation; Sect. 5 presents our main algorithm for surface reconstruction; Sect. 6 presents our results. A summary of this work is presented in Sect. 7.

2 Related Work

There is a vast amount of literature on 3D imaging. Readers are directed to Seitz et al. (2006) and Whöler (2009) for broad overviews; here only representative examples are discussed.

3D imaging has been approached using passive triangulation methods such as conventional stereo (e.g. Scharstein and Szeliski 2002), passive photometric methods such as shape from shading (e.g. Horn and Brooks 1989), active triangulation methods such as structured-light (e.g. Scharstein and Szeliski 2003) and active photometric methods such as photometric stereo (e.g. Woodham 1980). Hybrid methods that integrate two or more methods include approaches that combine shape from motion and photometric stereo (e.g. Higo et al. 2009), positional (3D points) data and normals (e.g. Terzopoulos 1988; Banerjee et al. 1992; Fua and Leclerc 1994; Lange 1999; Ikeuchi 1987; Bernardini et al. 2002; Chen et al. 2003; Nehab et al. 2005), visual hull and normals (Hernández et al. 2008),[1] and recently normals and volume carving (Vlasic et al. 2009).

All of the previously mentioned hybrid methods have the potential to be adapted to handle very high-resolution imagery as in our application. We opted for a solution that is closely related to the work by Nehab et al. (2005). The system presented in Nehab et al. (2005) used two cameras in a structured-light setup with one of the cameras also used to perform photometric stereo. Positional data and surface normals were fused using a linear formulation that resulted in a sparse-linear system. In their work, the surface geometry and the photometric data had approximately the same resolution. In our work, we have 100× more estimated surfaces normals than we do 3D points. At our resolutions (e.g. 5×5K surface normals), the sparse matrix proposed by Nehab et al. (2005) would have ∼150 million entries. Solving such a large matrix is not straight forward, even for out-of-core linear solvers such as Reid and Scott (2009). As such, we adopt a patch-wise strategy to the fusion process. In addition, to deal with the significant difference in resolutions, we use a multi-resolution pyramid approach to adaptively incorporate the geometric constraint from the low-resolution geometry during the surface integration.

In the case of the macro-lens imaging setup, our work intends to utilize auxiliary depth information obtained from depth-from-focus. These methods estimate an object's surface structure from two or more images with varying focus parameters. Notable examples include Darrell and Wohn (1988); Nayar and Nakagawa (1994); Xiong and Shafer (1993); Malik and Choi (2008). The basic idea involves determining when a point becomes in focus and relating that to the (calibrated) focal distance to the camera. While our approach draws on elements from prior techniques, as far as we are aware, the combination of photometric stereo and depth-from-focus with focal stacking is unique. In addition, our use of the focus images and photometric stereo lighting to improve the normal estimation and depth-from-focus results presents a new method for 3D imaging in restricted working environments.

3 System Setup

In this paper, we examine two ways of obtaining high-resolution surface normals and low-resolution depth using the following systems: (1) a large-format digital camera combined with structured-light; (2) a system consisting of a conventional DSLR and a macro lens. Both systems use four controllable lights and a mirrored sphere for light direction calibration. The photometric stereo technique in Woodham (1980) is used to obtain the surface normals. While we assume that the object is Lambertian, cross polarization is used to reduce specular reflection that arises when the object is not perfectly Lambertian. Polarization can effectively reduce most of the specular reflections from the objects. Our cross polarization is achieved by putting polarizers on the lens and in front of the four lights such that the polarizers are rotated to the angle that minimizes specular reflections (Nayar et al. 1997). The details of the two systems are described in the following.

3.1 Large-Format Digital Camera with Structured-Light

This system uses a large-format camera combined with a separate structured-light setup. Because the large-format camera used in the system requires roughly a minute to capture a single image, we opted to use an auxiliary video camera to perform the structured-light procedure. The two cameras and projector are calibrated by a physical calibration pattern. Figure 3a shows our setup.

Instead of using commercial large-format cameras (e.g. Anagramm And Digital Reproduction 1998), we use a custom-built 1.6gigapixel camera that uses a translation scanning back with an effective format of 450×300 mm. For more details of the large-format camera see http://dgcam.org. Figure 3b shows the scanning setup of an object that is roughly 20 cm in diameter. The resulting image of this object is ∼5×5K pixels.

The structured-light setup consists of a Benq MP624 projector and a 1,024 × 768 video camera. Standard binary gray-code patterns (Scharstein and Szeliski 2003) are used to estimate the low-resolution geometry. Figure 1b shows a small example of the 3D surface geometry estimated using the structured-light setup. There are slight pixelization-like artifacts due to inaccuracies in estimating the projected patterns' boundaries, however, since the low-resolution geometry serves only as a soft constraint in the surface reconstruction process our approach is insensitive to these errors.

3.2 DSLR with Macro Lens

Our macro-lens setup consists of a Canon EOS 1Ds Mark III camera and a Canon EF 50 mm f/2.5 Compact macro lens. In order to capture high-resolution data, we place the camera very close to the target object. As a result, we need to use focal stacking to capture our target objects. This also makes it hard to couple our system with structured-light setup. Figure 4 shows our setup. To capture the focal stack photometric stereo

[1] Also see http://carlos-hernandez.org/gallery/.

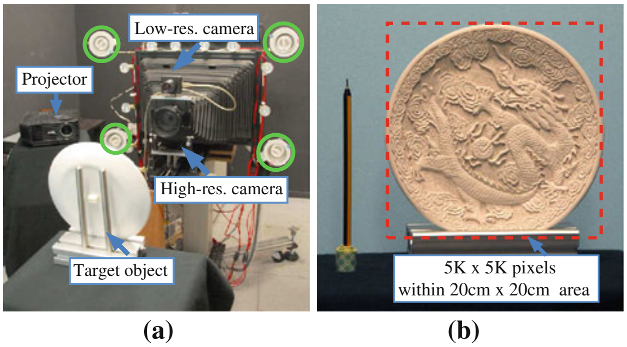

Fig. 3 Our first system: **a** The setup consists of an high-resolution large-format camera with four lights used for photometric stereo. A low-resolution video camera and digital light projector form the structured-light system. **b** The effective resolution of one of our objects is shown. Note the scale of the physical object, versus the pixel resolution. This results in a pixel resolution of over 600 samples per mm^2

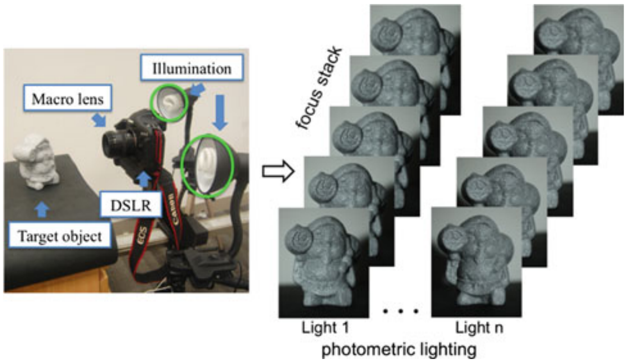

Fig. 4 Our second system: (*left*) The setup consists of a conventional DSLR with a macro lens and four lights used for photometric stereo. Note the two lights are shown for demonstration. In our experiments the lights are positioned much further from the object. (*right*) Both focal stack and photometric data are captured at the same time

data, we first focus on the nearest point on the target object. Then we increase the focus distance at every 3 mm. In the setup, we make sure that the depth-of-field for each image is about 3 mm so that each point on the object is in focus in at least one image. For each focus distance, the same four lights are used for photometric stereo.

4 Focal Stack Photometric Stereo

We first describe how we utilize focal stack data in the photometric stereo process as this is intended as a pre-processing step to our surface reconstruction algorithm shown in Sect. 5. We start with a short discussion on focal stack imaging and its relationship to surface normals. This is followed by a description of our normal estimation algorithm which globally optimizes constraints of defocus normals over the entire focal stack. Finally, a simple method to improve the depth-from-focus result using the available photometric stereo lighting is described.

4.1 Focal Stack and Normals

To understand the additional information pertaining to normals in the focal stack, let us first study the relationship of the surface normals when the captured images are out-of-focus. We assume the captured surface follows the Lambertian lighting model. For each point of the input image, we can represent the effect of defocus blur by the following convolution equation:

$$\begin{aligned} I(x, y) &= \sum_{(m,n)} I^*(x', y')K(m, n) \\ &= \sum_{(m,n)} \rho(x', y')|N^*(x', y') \cdot L(x', y')|K(m, n), \end{aligned} \tag{1}$$

where I is the captured defocus image, I^* is the ideal all-in-focus image, K is the spatially varying defocus blur kernel with its size proportionates to the depth of the scene, N^* is the surface normal, L is the lighting direction, ρ is the surface albedo, and $(x' = x\text{-}m,\ y' = y\text{-}n)$ is the local neighborhood of (x, y).

In photometric stereo, we assume that L is a directional light source and therefore constant within the same input image. Now, suppose for a local region, if ρ is constant, we can simplify Eq. (1) as follows:

$$I(x, y) = \rho L \cdot \left(\sum_{(m,n)} N^*(x', y')K(m, n) \right). \tag{2}$$

Solving Eq. (2) using standard least square methods (Woodham 1980), we obtain the normals which have undergone defocus blur:

$$\tilde{N}(x, y) = \sum_{(m,n)} N^*(x', y')K(m, n). \tag{3}$$

Note that although our formulation assumes that ρ is constant, in practice we find this is not necessary. This will be demonstrated in our results in Sect. 6.

4.2 Normals Refinement Using Deconvolution

Ideally, within the focal stack, there should be one normal estimation per pixel that is in-focus (or the best in focus). Hence, we could obtain all-in-focus normals by applying all-in-focus methods (e.g. Hausler 1972; Agarwala et al. 2004) over the normals at different focus distances. However, we found that computing an all-in-focus normal map this way resulted in a noisy result. This is due to the quantization effect in the focal stack and noise in estimating where a surface patch is in focus. Our aim is instead to estimate the normals globally using the entire focal stack.

For a focal stack with M number of levels, we can obtain M observations of normals as follows:

$$\begin{aligned}
\tilde{N}_1(x, y) &= \sum_{(m,n)} N^*(x', y')K_1(m, n) \\
\tilde{N}_2(x, y) &= \sum_{(m,n)} N^*(x', y')K_2(m, n) \\
&\vdots \\
\tilde{N}_M(x, y) &= \sum_{(m,n)} N^*(x', y')K_M(m, n).
\end{aligned} \tag{4}$$

Solving each individual equation alone in Eq. (4) is well-known to be an ill-posed problem. However, when we combine all the equations together, this problem becomes well-posed as shown in Agrawal et al (2009). This can be formulated into a set of linear equations by rewriting Eq. (4) into $AN^* = b$ with:

$$\begin{aligned}
A &= \begin{bmatrix} K_1 \\ \vdots \\ K_M \end{bmatrix}^T \begin{bmatrix} K_1 \\ \vdots \\ K_M \end{bmatrix} + w\left(G_x^T G_x + G_y^T G_y\right) \\
b &= \begin{bmatrix} K_1 \\ \vdots \\ K_M \end{bmatrix}^T \begin{bmatrix} \tilde{N}_1 \\ \vdots \\ \tilde{N}_M \end{bmatrix},
\end{aligned} \tag{5}$$

where G_x and G_y are the $x-$ and $y-$ derivative filters used as regularization to help suppress noise and ringing artifacts in the deconvolution. The term w is the regularization weight.

Since the defocus kernel is different at each pixel, the deconvolution process is performed for each pixel individually. In our implementation, we calibrated the defocus kernel K using a textured pattern for each level of defocus in the focal stack. Note that if the lens optics are known, the K can be computed directly. The defocus kernel is selected according to the depth map estimated from the depth-from-focus which will be detailed in next subsection. While the estimated defocus kernel might be inaccurate, the multiple observations of the blurry normals and the neighborhood regularization in Eq. (5) help ameliorate these estimation errors. Figure 5 shows the comparisons between our estimated normals and the all-in-focus normals obtained using an all-in-focus method (Hausler 1972). See Sect. 6.1.1 for the details of how we compute the normals using this method. While these look similar, on careful inspection it is clear that the all-in-focus normal map (Fig. 5b) is more noisy than that obtained with regularized normal map (Fig. 5a). Further quantitative evaluations of our normal estimation method are given in Sect. 6.

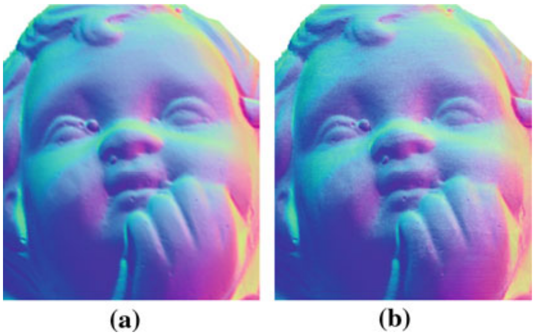

Fig. 5 The estimated normals, with **a** and without **b**, deconvolution refinement. While similar at first glance, on careful inspection it is apparent that the normal map in (**b**) exhibits more noise than the normals in (**a**)

4.3 Depth-from-Focus Exploiting Photometric Lighting

Techniques for depth-from-focus return a depth map from the focal stack via edges/textures sharpness analysis. These measurements depend greatly on the rich texture information on the object surface. In situations where the object's surface does not contain any texture, the above measurements may fail since there is no obvious difference between the in-focus image and the out-of-focus images in the focal stack.

In photometric stereo, scenes are captured with varying illuminations under different lighting directions. The controlled light sources produce shadings or shadows according to the geometry and curvature of object surface, regardless of local albedo. In other words, the shadings/shadows can still exist when the local albedo is the same. With the induced discontinuities from shadings or shadows, we can analyze the amount of defocus even when the object is homogeneous in color. Note that when more lighting directions are used in photometric stereo, we can get more accurate depth since we have more observations of shading/shadows discontinuities.

We take the photometric stereo images at each focus distance as a trade off of additional capturing time. The lighting directions are fixed and calibrated so that at each focus distance, we can find the corresponding photometric stereo images in the focal stack for focus analysis. The pseudo code of our depth-from-focus algorithm is detailed in Algorithm 1.

Algorithm 1 Depth-from-Focus Exploiting Photometric Lighting

```
$F_{max} = 0$
$\bar{d} = 0$
for each illumination $i$ do
  $F^i_{max} = 0$
  $d^i = 0$
  for each focus distance $k$ do
    Compute focus measure $F^i_k$
    if $F^i_{max} < F^i_k$ then
      $F^i_{max} = F^i_k$
      $d^i = d_k$
    end if
  end for
  if $F_{max} < F^i_{max}$ then
    $F_{max} = F^i_{max}$
    $\bar{d} = d^i$
  end if
end for
```

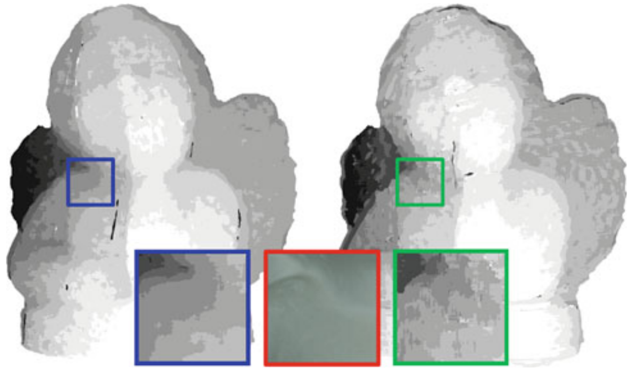

Fig. 6 The estimated depth map using depth-from-focus, with (*left*) and without (*right*) photometric lighting. The zoomed-region shows that the estimated depth is less noisy with photometric lighting. Note that the *red box* in the middle shows the zoomed-region of an input image

The term F_k^i is the focus measured under lighting direction i at the focus distance k. We use the Sum-Modified-Laplacian introduced by Nayar and Nakagawa (1994) for focus analysis. The term d_k is the depth corresponding to the focus distance k. The term $\bar{d}$ is our resulting depth estimation. In addition to the depth map, we will also compute a confidence map C which measures the reliability of the depth value computed at each pixel location. Our confidence map is computed using the following equation:

$$C(x, y) = \frac{\sum_{(j \in \mathscr{J})} F_{max}^j(x, y)}{\sum_{(i \in \mathscr{L})} F_{max}^i(x, y),} \tag{6}$$

where $\mathscr{J} = \{i | d^i = \bar{d}\}$, $\mathscr{L}$ is the set of all illuminations, $\bar{d}$ is the overall depth estimate and d^i is depth estimate for ith illumination, as computed in Algorithm 1. Figure 6 shows the comparison between our estimated depth using photometric lighting and usual lighting. The zoomed-in region shows the estimated depth is less noisy when the photometric lighting is used. Under ideal conditions, depth-from-focus produces the same resolution depth map as the input images. In practice, however, the depth maps are often noisy even with our photometric lighting. We further reduce the effects of noise by removing the low confidence depths using C and a median filtering followed by downsampling the depth map by a factor of four. We use this sparse/low-resolution depth map as the low-resolution constraint in our surface reconstruction algorithm (see Sect. 5).

5 Surface Reconstruction Algorithm

This section describes the surface reconstruction algorithm of our framework. The basic algorithm to reconstruct a surface from normals is described first. This is followed by a description on how to include the low-resolution geometry constraint and boundary connectivity constraint into the algorithm. Finally, the steps of the multi-resolution strategy is described in detail.

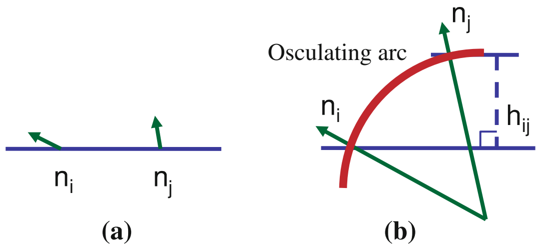

Fig. 7 The osculating arc constraint (Wu et al. 2008) for surface reconstruction. Given the normal configuration $\{n_i, n_j\}$ between neighborhood pixel i and j in **a**, we can uniquely define the relative height h_{ij} in **b** by using an osculating arc to connect n_i and n_j with minimum curvature

5.1 Surface from Normals

Given a dense set of normals the goal is to reconstruct a surface that satisfies the normals' orientation constraint. We use the recent approach presented by Wu et al. (2008) for obtaining a surface from normals that constrains the estimated surface using an osculating arc between neighboring normals (see Fig. 7). This problem can be cast as a least-square problem that minimizes the following energy function:

$$E(S|\vec{n}) = \sum_{i}^{N} \sum_{j \in \mathscr{N}(i)} \left((S_i - S_j) - h_{ij}\right)^2, \tag{7}$$

where S is the surface we want to reconstruct, $(S_i - S_j)$ is the first derivative of S in discrete form, h_{ij} is the relative height defined by the osculating arc constraint between neighborhood pixels, $\mathscr{N}(i)$ is the first order neighborhood of a pixel, and N is number of pixels. A qualitative comparison of the osculating arc constraint with other surface from normals algorithms can be found in Wu et al. (2008). While the osculating arc constraint does not explicitly deal with depth discontinuity, Wu et al. (2008) showed that the osculating arc constraint produces the least distorted results compared with methods that explicitly consider discontinuity such as affine transformation and M-Estimator (Agrawal et al. 2006).

Equation (7) can be solved using Gauss-Seidel iteration. At each iteration, the surface height is updated according to the following equations:

$$S_i^{t+1} = S_i^t + \lambda_1 \xi_1$$
$$\xi_1 = \frac{1}{|\mathscr{N}(i)|} \sum_{j \in \mathscr{N}(i)} (h_{ij} - (S_i^t - S_j^t)), \tag{8}$$

where $|\mathscr{N}(i)|$ is the number of neighborhood pixels, $\lambda_1 = 0.9$ is the step size and t is the iteration index. Note that h_{ij} is the same for all iterations and can be pre-computed.

5.2 Low-Resolution Geometry Constraint

Because photometric stereo inherently captures only local reflection information rather than global structure, many surface from normal reconstruction approaches do not accurately reflect the real surface geometry. As discussed in Sect. 2, one strategy to overcome this is to incorporate positional information in the reconstruction process.

Our low-resolution geometry constraint is modeled using the following equation:

$$E(S|L) = \sum_{i}^{M} \left(|d(h(S_i)) - L_i| - \Delta\right)^2 , \tag{9}$$

where L is the low-resolution geometry, M is the number of pixels in the low-resolution geometry, $h(\cdot)$ is a Gaussian convolution process with radius equals to two times the downsample rate, $d(\cdot)$ is a downsample operation to match the high-resolution normals to the low-resolution geometry, and $|\cdot|$ is the L1 norm (absolute value) of the errors. The term Δ is a parameter controlling the amount of depth tolerance for surface details to be reconstructed and refined by the normals.

With the additional low-resolution geometry constraint, the iterative update equation in Eq. (8) is updated as follows:

$$\begin{aligned} S_i^{t+1} &= S_i^t + \lambda_1 \xi_1 + \lambda_2 \xi_2 \\ \xi_2 &= \begin{cases} h(u(L_i - d(h(S_i)))), & \text{if } |d(h(S_i)) - L_i| > \Delta \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \tag{10}$$

where $u(\cdot)$ is an upsample operator. The effect of our low-resolution geometry constraint is shown in Fig. 8. The value of Δ is estimated according to the variance of surface details reconstructed from normals and can be spatially varying.

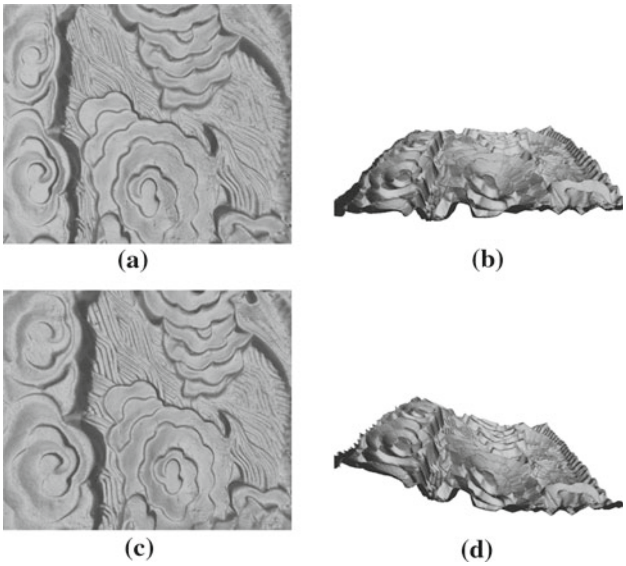

Fig. 8 Example of surface reconstruction with/without including the low-resolution geometry constraint: **a** reconstructed surface from normals only, **b** a side view of (**a**), **c** reconstructed surface with low-resolution geometry constraint, **d** a side view of (**c**). Note that this patch is at the edge (bending) of a plate. See Fig. 19 for the overall geometry and surface normals of the plate

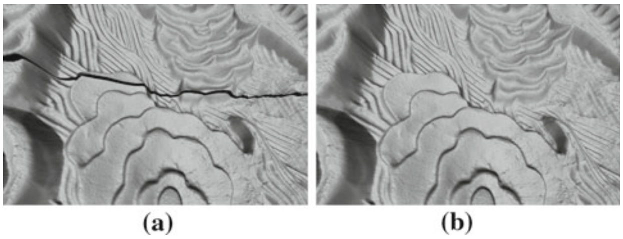

Fig. 9 Effect of the boundary connectivity constraint: **a** without the boundary connectivity constraint, and **b** with the boundary connectivity constraint

5.3 Boundary Connectivity Constraint

As discussed in Sect. 2, the high-resolution of the photometric stereo component makes it challenging to perform integration on the entire surface in one pass. To overcome this the surface can be subdivided into more manageable sized patches and each patch is reconstructed individually. This leads to a problem that the boundaries of adjacent patches may not be properly aligned after reconstruction. To overcome this, we add a boundary connectivity constraint described by the following equation:

$$E(S|B) = \sum_{i \in \Omega} (S_i - B_i)^2, \tag{11}$$

where Ω is the overlapping area of neighborhood surface patch, B is a surface computed by blending the intermediate reconstructed surface in Ω between neighborhood patches using linear feathering. Adding the boundary constraint into Eq. (8), we get:

$$\begin{aligned} S_i^{t+1} &= S_i^t + \lambda_1 \xi_1 + \lambda_2 \xi_2 + \lambda_3 \xi_3 \\ \xi_3 &= \begin{cases} B_i - S_i, & \text{if } i \in \Omega \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{12}$$

For this boundary connectivity constraint, the weight of λ_3 during the update iterations needs to be adjusted as the system is iterated. In the initial estimation, λ_3 is equal to zero, and its weight is gradually increased as the number of iterations increases. This allows the surface patch to be reconstructed freely at initial iterations and later refined to meet the boundary of neighborhood patches. In our implementation, λ_3 and B are updated at every 100 iterations. With this boundary connectivity constraint, surface reconstruction can be done in parallel and the problem of resolution is no longer an issue. The effect of this boundary constraint is shown in Fig. 9. For the results in this paper, surface patches are taken to be of size 1024 × 768 with overlaps of 100 pixels (i.e. 10 %).

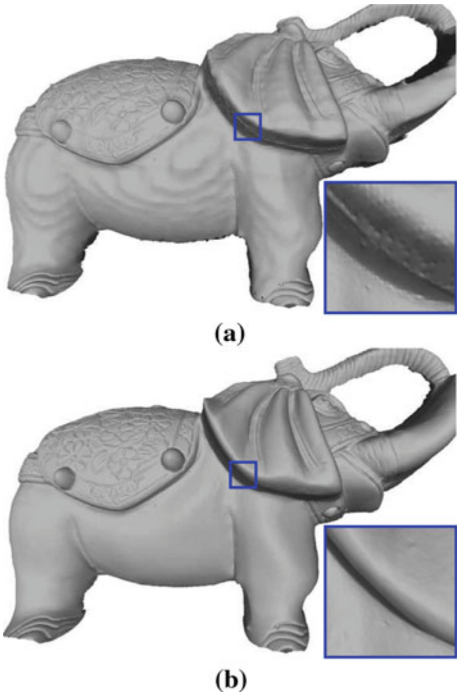

Fig. 10 Example of the benefits of the multi-resolution scheme: **a** our surface reconstruction directly using the low-resolution geometry, **b** reconstructed surface using the multi-resolution scheme. The 3D surface in **a** shows noticeable quantization errors due to the low-resolution geometry

5.4 Multi-Resolution Pyramid Approach

Due to the very large differences in resolutions between the surface normals and the low-resolution geometry, directly adding the surface normals to the low-resolution geometry will result in noisy reconstruction as shown in Fig. 10. To avoid this, our surface reconstruction is done in a multi-resolution pyramid fashion. The main purpose of using the pyramid approach is to correct the low-resolution geometry using normals at the equivalent level before we use the geometry as soft constraint at a higher resolution. In the case of the large-format camera setup, the multi-resolution pyramid approach also allows us to resolve small misalignments between the high-resolution normals and low-resolution geometry due to device calibration errors.

We divide the pyramid uniformly into different levels starting at the resolution used to capture the low-resolution geometry (i.e. 1,024 × 768 for the large-format camera setup and 1,404 × 936 for macro-lens setup). For each level, instead of downsampling the estimated high-resolution normals, we downsampled the high-resolution input images and estimate the normals from the dowsampled images. We run our surface reconstruction algorithm described in Eq. (12) with the results from previous level as the low-resolution constraint. For the lowest resolution, the low-resolution geometry estimated by structured-light is used. Figure 11 shows our intermediate surface reconstruction results (i.e. the evolution) at different levels in the pyramid.

6 Results

In this section we first demonstrate our normal estimation refinement using focal stacking on both synthetic and real objects. This is followed by our high-resolution results captured by our large-format camera with structured-light.

6.1 Photometric Stereo using Focal Stacking

We test our algorithm on synthetic and real objects. Since ground truth normals are difficult to obtain for this type of setup, we performed a series of synthetic experiments using the *Maya* rendering software to simulate shallow depth-of-field imaging with 3D models from which we can compute normals for ground truth. Experiments are also performed on real objects which show the difference in quality using both our estimated normals and depth-from-focus algorithms.

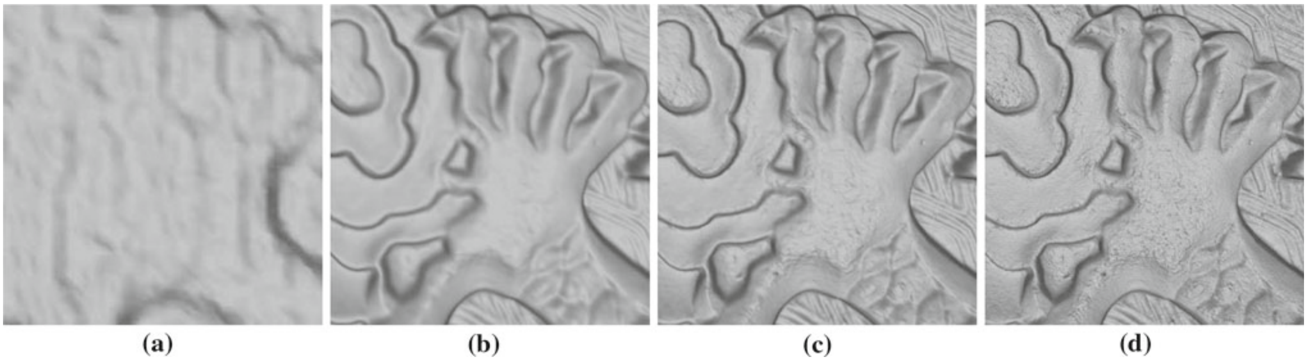

Fig. 11 Evolution of our 3D surface up the multi-resolution pyramid: **a** low-resolution geometry, **b** intermediate result at the lowest level of the pyramid, **c** the third level, **d** the last level and final 3D reconstruction

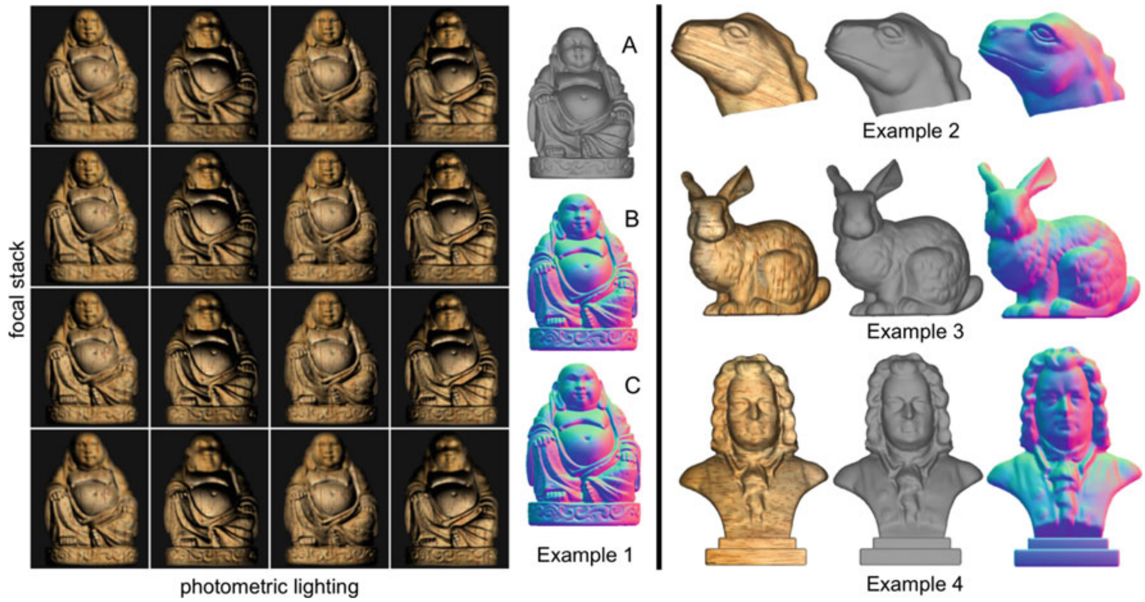

Fig. 12 On the left, we show the full input images of *example 1*. The object without texture (**A**), normal map of the ground truth (**B**) and our method (**C**) are shown too. On the right, *example 2, 3* and *4* are shown with and without texture, as well as normal maps computed from our method

6.1.1 Synthetic Examples with Ground Truth

Generating Synthetic Data We use four objects with known geometry and surface normals. We select the object material to be purely Lambertian. To unify the camera setup in *Maya*, all the objects are scaled to the similar size. To simulate the focal stack, we set the camera in *Maya* at a fix distance looking at the object. We turn on *Maya*'s depth-of-field setting and set the camera aperture to $f1.0$ to simulate a shallow depth-of-field at each focus distance. We then render the objects using this synthetic camera model. In order to test the effect of texture on our method, we render each object both with and without texture.

The camera in *Maya* is focused at four different distances from the closest point of the object to the furthest point observable from the camera. To simulate photometric stereo, we set up four different directional lights. The intensities of all the lights are set to the same value. At each focus distance, four images are generated such that each image is rendered with only one light on. In total, 16 images are rendered and serve as input for our methods.

Calibrating Defocus Kernels To perform our normal refinement, we need to calibrate the defocus kernels for each focus distance. While we could compute this from the lens model used by *Maya*, to provide a more realistic experiment, the defocus kernels are estimated in the same manner as with real objects. This is done by placing a textured plane at the nearest focus distance. Then we render multiple images with the camera focused at each focus distance. As a result, the textured plane is blurred with different defocus kernels at different focus distance. These images are then used for defocus kernel estimation.

Comparisons and Results With the focal stack photometric stereo images rendered from *Maya*, we compute the normals using our method described in Sect. 4. As a comparison, we also compute all-in-focus normals using two focal stack methods: a classic all-in-focus method (Hausler 1972) and a recent method based on graph-cut (Agarwala et al. 2004). For the latter, we use the code provided by the authors' webpage. We first compute the normals at the different focus distances. These normals are then used as input (instead of RGB images) to the two methods for computing all-in-focus normals. Figure 12 shows our synthetic examples with and without texture, as well as normals computed using our method. In Table 1, we show the mean angular errors (in degrees) between normals computed by ours and the two all-in-focus methods and the ground truth, for each object used. To show our method is not just a matter of filtering, we applied bilateral filtering on normals computed by the two all-in-focus methods. While the overall error is relatively small for the all-in-focus approaches, our approach is consistently better. We also applied our method on the four examples without texture (see the last two rows of Table 1) with virtually the same results. This shows the ability of our approach to handle textured and textureless inputs. Because of the photometric lighting, our depth-from-focus gives better estimations in regions with less texture. Using these depth estimation to assist in selecting the correct defocus kernel, our normal refinement technique achieves comparable results even for textureless objects.

Table 1 Comparison on average angular error (in degrees) of normals among our method and the all-in-focus methods with and without bilateral filtering

Textured	Ex1	Ex2	Ex3	Ex4
Hausler (1972)	2.51	1.07	0.97	2.12
Agarwala et al. (2004)	2.44	0.88	1.07	1.63
Hausler (1972) + BL	2.71	1.12	1.10	2.22
Agarwala et al. (2004) + BL	2.12	0.88	0.97	1.50
Ours	2.06	0.77	0.69	1.41
Ours textureless	2.05	0.78	0.70	1.41

Comparisons are on textured objects. The last row shows the results are virtually identical when the object is textureless

6.1.2 Real Objects

This section shows several real objects captured by our DSLR with macro lens setup described in Sect. 3.2. For each object, normals and depths were estimated using the methods described in Sect. 4. Then the surfaces were reconstructed using the technique in Sect. 5. Each surface was generated using 800 iterations (per patch) of our surface reconstruction algorithm with the boundary constraint applied once after every 100 steps. Each of the objects required about 12 patches. Similar to the synthetic experiments, we precalibrated the defocus kernels using a patterned board.

In Fig. 13, we show an example with heavy texture and pitted surface. The zoomed-region shows the difference of normals and relighted images from our method and all-in-focus method (Hausler 1972). Figures 14 and 15 show estimated normals and 3D reconstructions of a *statue* and an *angel* figurine. The zoomed-regions on the right show the comparisons of our method and all-in-focus methods (Hausler 1972; Agarwala et al. 2004). Figure 16 shows the estimated normals and the 3D reconstruction of a *duck* figurine. The zoomed-regions on the right show the comparisons of our methods with and without normal refinement, and depth-from-focus with and without exploiting photometric lighting. The improvements in the results using the depth-from-focus algorithm are clear, producing less noticeable artifacts in the coarse depth estimation. The normal map improvements are more subtle, but on close inspect reveal that our approach has less noisy normals, resulting in smoother results that still contain small details present on the objects surface. For the examples produced using Agarwala et al. (2004), seams are often noticeable in the final results due to the graph-cut algorithm.

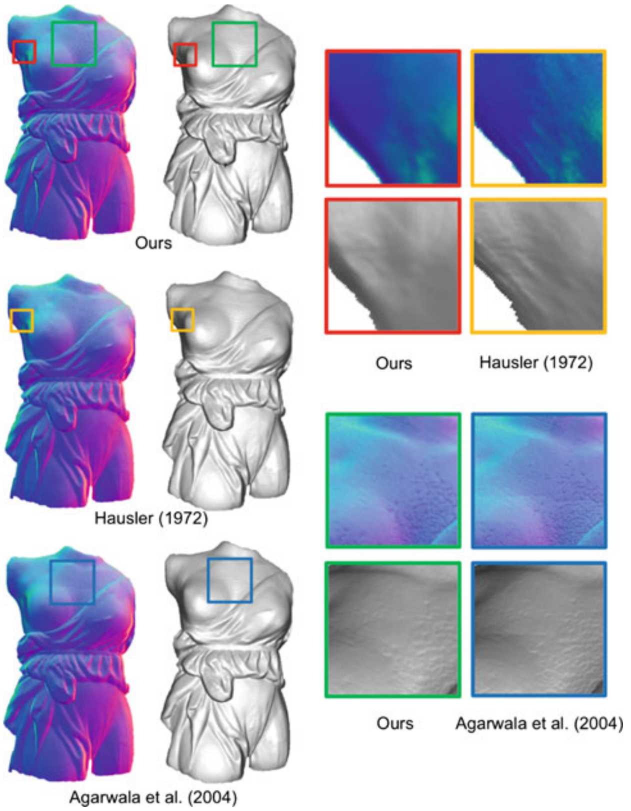

Fig. 14 Normal map and 3D reconstruction result of *statue* figurine. (*Left*) normal map and 3D reconstruction of our approach, Hausler (1972) and Agarwala et al. (2004). (*Top right*) the comparison of a zoomed-region of our and Hausler (1972). (*Bottom right*) the comparison of a zoomed-region of our and Agarwala et al. (2004)

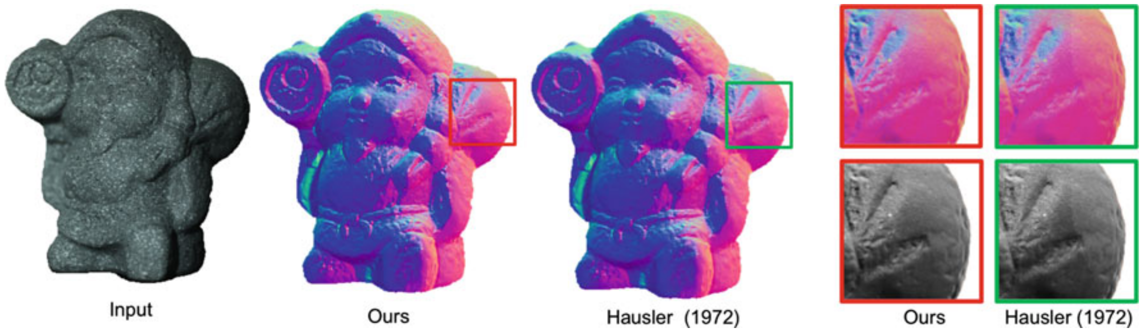

Fig. 13 An example with heavy texture and pitted surface. (*Left*) an example of input image, normal map computed by our method and Hausler (1972). (*Right*) the comparison of a zoomed-region of our and Hausler (1972). Note that *bottom right* shows re-lighted images from normals

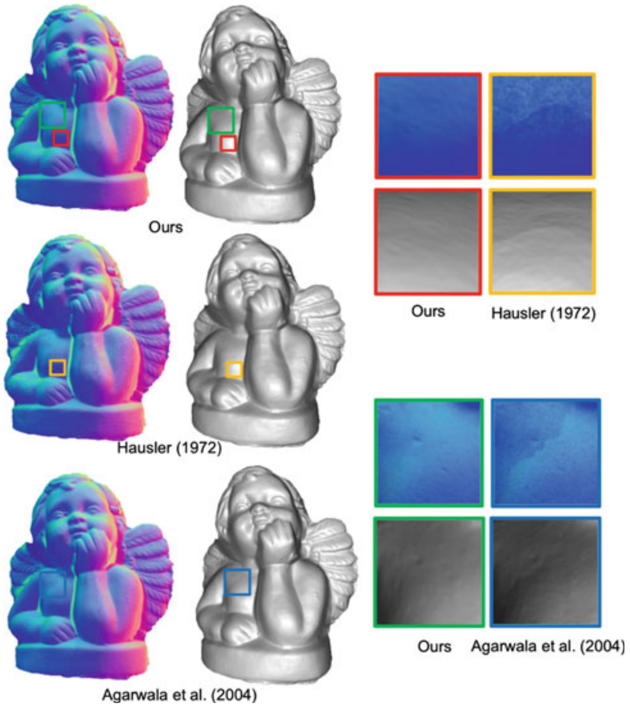

Fig. 15 Normal map and 3D reconstruction result of *angel* figurine. (*Left*) normal map and 3D reconstruction of our approach, Hausler (1972) and Agarwala et al. (2004). (*Top right*) the comparison of a zoomed-region of our and Hausler (1972). (*Bottom right*) The comparison of a zoomed-region of our and Agarwala et al. (2004)

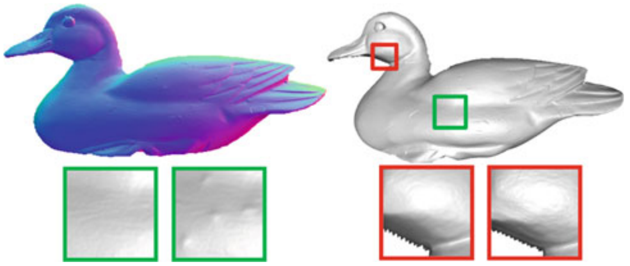

Fig. 16 Normal map and 3D reconstruction result of *duck* figurine. (*Bottom left*) the comparison of a zoomed-region with (*left*) and without (*right*) photometric depth-from-focus. (*Bottom right*) the comparison of a zoomed-region with (*left*) and without (*right*) normal refinement

6.2 Results from Our Large-Format Camera System

This section shows several results captured by our large-format camera system described in Sect. 3.1. The number of iterations (per patch) and steps for boundary constraint are the same as those used in Sect. 6.1.2. Besides qualitative results, we also evaluate our reconstruction results quantitatively for one of the objects.

Figure 17 shows 3D reconstruction of an *elephant* figurine which is ∼15 cm wide. Figure 18 shows example of a *man* figurine of roughly 12 cm high. The objects required 9 patches and resulted in about 6.5 million reconstructed 3D points, while the *man* required 12 patches and resulted in about 4.5 million reconstructed 3D points. Both of these results show exceptional surface detail. The *man* figurine is further zoomed to reveal detail that would require a magnifying glass to be seen (properly) with the unaided eye. For visual comparison, zoom and double-zoom regions from one of the input images are also shown.

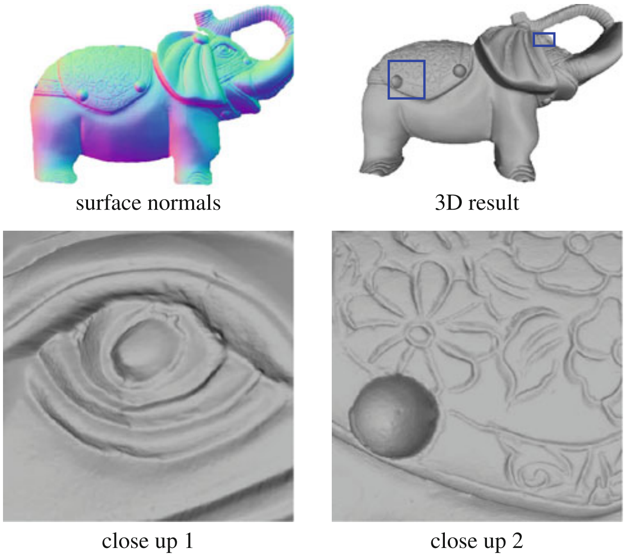

Fig. 17 3D reconstruction of the *elephant* figurine. The zooms show exceptional detail on the surface of the object

Finally, we compare our result, a scanned *dragon plate*, with that obtained from an industrial standard high end laser scanner in Fig. 19. We took the object to an industrial scanning facility using a Konica Minolta Range 7. This serves as our baseline comparison against a state-of-the-art industrial laser scanner. The finest scanning resolution that can be obtained by the laser scanner is 168 samples per mm^2, while our sampling rate is 600 samples per mm^2. The plate required 25 patches and resulted in about 21.5 million reconstructed 3D points. The state-of-the-art scanner reports to have a scanning accuracy of 40 microns. We can see that on the double-zoom of these two surfaces, we reveal detail while the result from the laser scanner is almost completely flat. Zoomed regions from one of the input images are also provided to aid the visual comparison. Note that in order to capture the whole plate by the laser scanner, several scans were performed and stitched together. Our approach, on the other hand, was able to image the entire 3D object in one pass. In addition, even though we use a patch-wise approach in surface integration, our surface does not contain any blocking or pixelization artifacts. This helps demonstrate the effectiveness of our boundary connectivity constraint and multi-resolution pyramid approach. Comparing the surface depth, our estimated surface depths are consistent with surface depths captured by laser scanner.

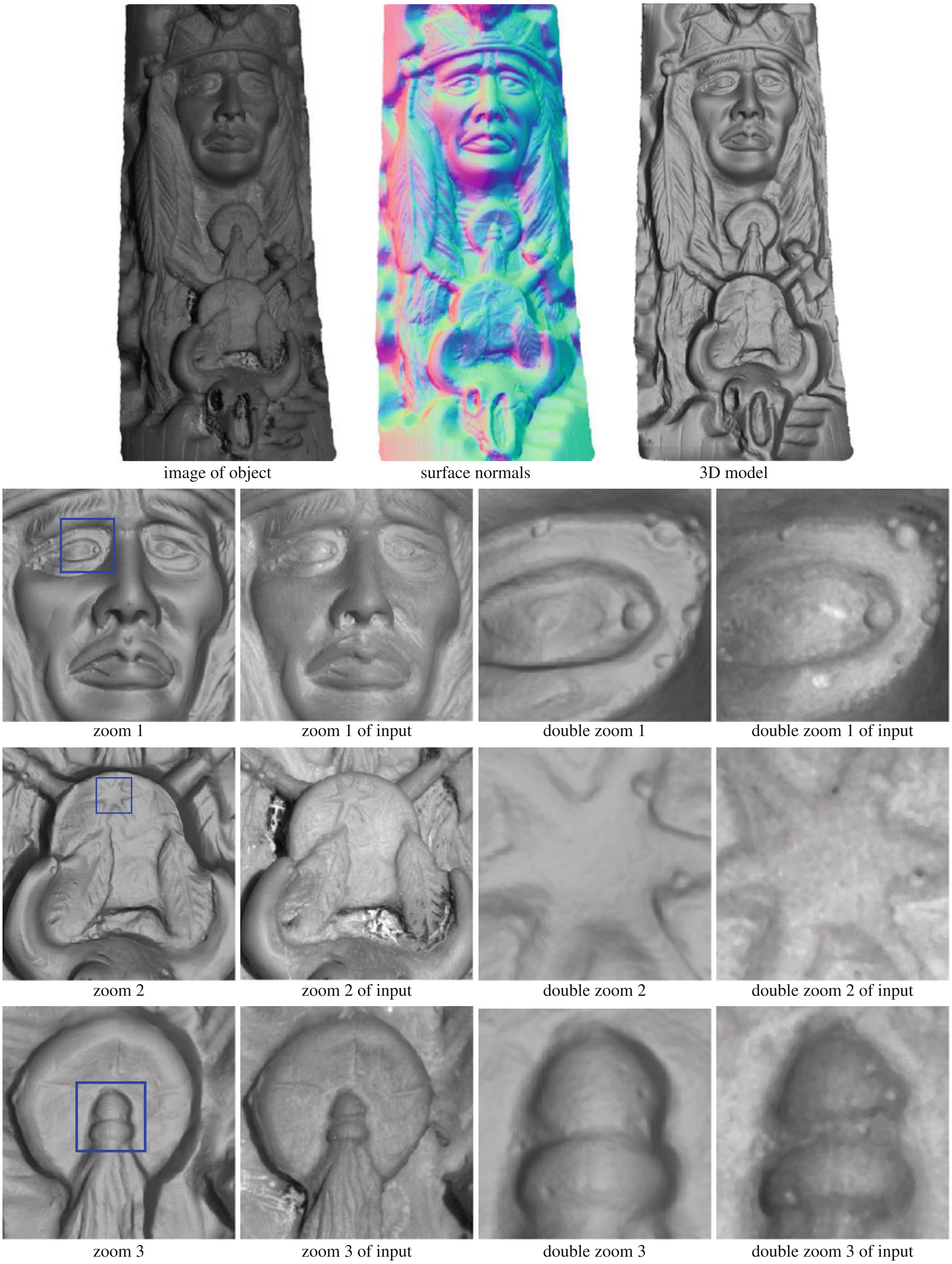

image of object

surface normals

3D model

zoom 1

zoom 1 of input

double zoom 1

double zoom 1 of input

zoom 2

zoom 2 of input

double zoom 2

double zoom 2 of input

zoom 3

zoom 3 of input

double zoom 3

double zoom 3 of input

Fig. 18 3D reconstruction of the *man* figurine. Due to the high-resolution of the 3D scan, we can show a zoom and “double zoom” of the 3D surface. We also show the zoom and “double zoom” from one of the input images. This double zoom reveals detail that would require a magnifying glass to see properly

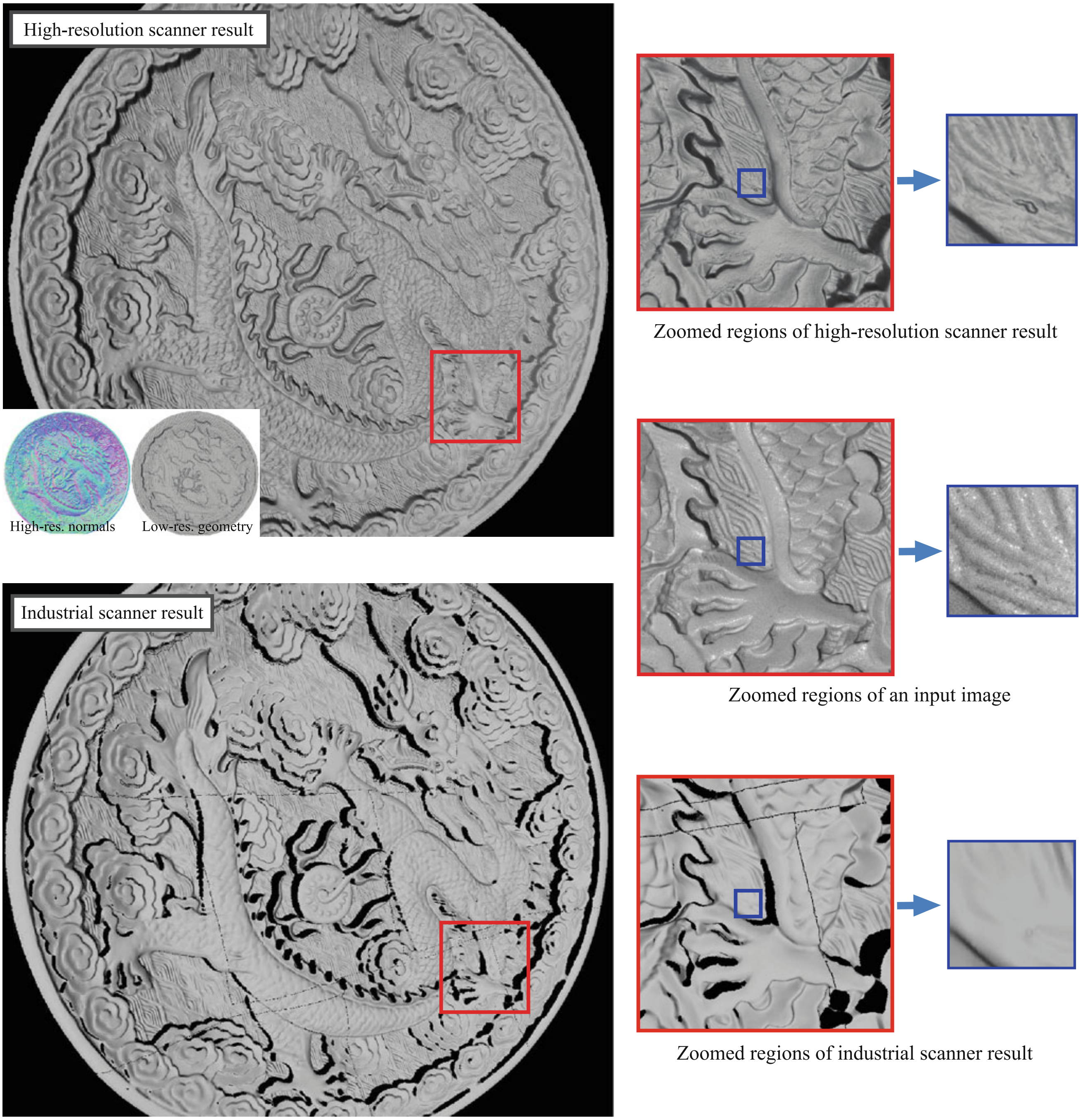
High-resolution scanner result
High-res. normals
Low-res. geometry
Industrial scanner result
Zoomed regions of high-resolution scanner result
Zoomed regions of an input image
Zoomed regions of industrial scanner result

Fig. 19 Full-size comparison with an industrial laser scanner. Shown are the full 3D reconstruction from our approach and that from a Konica Minolta Range 7 industrial scanner. Insets for our approach show the surface normals and low-resolution geometry. Zoomed and double zoomed regions from both 3D results and one of the input images show that while the two scans reveal that our result contains considerable more surface detail, both appear to reflect the correct geometry. Note that the Konica Minolta Range 7 specifications state a scanning accuracy of up to for ±40 mu

To quantitatively evaluate the effectiveness of our low-resolution geometry constraint in assisting the 3D reconstruction, we compute the distances on the *dragon plate* example from the surface captured with the laser scanner to our reconstructed surface with and without the low-resolution geometry constraint. Due to the large resolution difference, we have to apply preprocessing steps to make the evaluation feasible and meaningful. We first downsampled our

Table 2 Numerical evaluation on the reconstructed *dragon plate* surfaces

Distance to laser scanner result	With low-res	Without low-res
Max	0.01954	0.03328
Mean	0.00167	0.00803
RMS	0.00299	0.01224

We use Metro (Cignoni et al. 1998) to compute the max, mean and RMS distances from the laser scanner result to the surfaces reconstructed with/without the low-resolution constraint. The reported numbers are with respect to the bounding box diagonal

high-resolution surfaces into the same level as the laser scanner result. Then a low-pass filter was applied to the downsampled surfaces. After alignment of the surfaces (manual alignment followed by automatic alignment using ICP), we used *Metro* (Cignoni et al. 1998) to compute the mean, max, and root mean square (RMS) distances of the surfaces. Table 2 shows the *Metro* output. While this evaluation is only an approximation due to the additional errors introduced by down-sampling, filtering, and alignment, we can nonetheless see our surface reconstructed with the low-resolution constraint is more consistent to the laser scanner result than the one reconstructed without the low-resolution constraint.

7 Discussion and Summary

This paper describes a 3D imaging framework that combines high-resolution photometric stereo data and low-resolution depth. Through either a macro-lens setup or large-format camera setup, our system is able to capture 3D surfaces at more than 600 samples per mm^2. Based on these imaging scenarios, we first show how to use the focal stack data in the photometric stereo process by introducing a method to regularize normals against the varied focused images to improve normal estimation. We then propose a multi-resolution patch-based approach that combines surface normals and depth samples with vast resolution differences. Our results demonstrate some of the most detailed 3D imaging data captured to date.

We note that the work presented in this paper operates within a conventional photometric stereo context which has many known issues that remain unsolved (e.g Lambertian surface assumption, albedo estimation, use of light sources and their calibration, etc). Another issue is sharp depth discontinuities that can cause distortion in the final reconstruction. One method to help overcome this is to include a discontinuity map (Wu and Tang 2006) to help constrain the surface reconstruction within continuous regions. This is an interesting area that warrants further research effort. In addition, our future work seeks to examine if information offered by the very high-resolution imagery or focal stack data may help overcome some of these limitations, especially in estimating more complex lighting models. Another avenue for future work is to attempt to derive second-order surface geometry information from the blur profiles of surface points that are presented in the focal stack data. Such information may be useful in further improving the surface normal estimation.

References

Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S., Colburn, A., Curless, B., et al. (2004). Interactive digital photomontage. *ACM Transactions on Graphics (SIGGRAPH)*, *23*(3), 294–302.

Agrawal, A., Raskar, R., & Chellappa, R. (2006). What is the range of surface reconstructions from a gradient field? In *European conference on computer vision (ECCV)*. Graz, Austria: Springer.

Agrawal, A., Xu, Y., & Raskar, R. (2009). Invertible motion blur in video. *ACM Transactions on Graphics (SIGGRAPH)*, *28*(3), 1–8.

Anagramm and Digital Reproduction (1998). http://www.linhofstudio.com. Accesed 1 August 2011.

Banerjee, S., Sastry, P., & Venkatesh, Y. (1992). Surface reconstruction from disparate shading: An integration of shape-from-shading and stereopsis. In *11th IAPR International conference on pattern recognition*. The Hague, The Netherlands: IEEE Computer Society.

Bernardini, F., Rushmeier, H., Martin, I. M., Mittleman, J., & Taubin, G. (2002). Building a digital model of Michelangelo's Florentine Pieta. *IEEE Computer Graphics and Applications*, *22*(1), 59–67.

Chen, C. Y., Klette, R., & Chen, C. F. (2003). *Shape* from photometric stereo and contours. In *Proceedings of computer analysis of images patterns (CAIP)*. Groningen, The Netherlands: Springer.

Cignoni, P., Rocchini, C., & Scopigno, R. (1998). Metro: Measuring error on simplified surfaces. *Computer Graphics Forum*, *17*(2), 167–174.

Darrell, T., & Wohn, K. (1988). Pyramid based depth from focus. In *Computer vision and pattern recognition (CVPR)*. Ann Arbor, MI: IEEE Computer Society.

Fua, P., & Leclerc, Y. G. (1994). Using 3-dimensional meshes to combine image-based and geometry-based constraints. In *European conference on computer vision (ECCV)*. Stockholm, Sweden: Springer.

Hausler, G. (1972). A method to increase the depth of focus by two step image processing. *Optics Communications*, *6*(1), 38–42.

Hernández, C., Vogiatzis, G., & Cipolla, R. (2008). Multi-view photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(3), 548–554.

Higo, T., Matsushita, Y., Joshi, N., & Ikeuchi, K. (2009). A hand-held photometric stereo camera for 3D modeling. In *Computer vision and pattern recognition (CVPR)*. Miami, FL: IEEE Computer Society.

Horn, B., & Brooks, M. (1989). *Shape from shading*. Cambridge: MIT Press.

Ikeuchi, K. (1987). Determining a depth map using a dual photometric stereo. *International Journal of Robotics Research*, *6*(1), 15–31.

Lange, H. (1999). Advances in the cooperation of shape from shading and stereo vision. In *Proceedings 3DIM*. Ottawa, Canada: IEEE Computer Society.

Lu, Z., Tai, Y. W., Ben-Ezra, M., & Brown, M. S. (2010). A framework for ultra high resolution 3D imaging. In *Computer vision and pattern recognition (CVPR)*. San Francisco, CA: IEEE Computer Society.

Malik, A. S., & Choi, T. S. (2008). A novel algorithm for estimation of depth map using image focus for 3d shape recovery in the presence of noise. *Pattern Recognition*, *41*(7), 2200–2225.

Nayar, S. K., & Nakagawa, Y. (1994). Shape from focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *16*(8), 824–831.

Nayar, S. K., Fang, X. S., & Boult, T. (1997). Separation of reflection components using color and polarization. *International Journal of Computer Vision*, *21*(3), 163–186.

Nehab, D., Rusinkiewicz, S., Davis, J., & Ramamoorthi, R. (2005). Efficiently combining positions and normals for precise 3d geometry. *ACM Transactions on Graphics (SIGGRAPH)*, *24*(3), 536–543.

Reid, J. K., & Scott, J. A. (2009). An out-of-core sparse cholesky solver. *ACM Transactions on Mathematical Software*, *36*(2), 1–33.

Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *47*(1–3), 7–42.

Scharstein, D., & Szeliski, R. (2003). High-accuracy stereo depth maps using structured light. In *Computer vision and pattern recognition (CVPR)*. Madison, WI: IEEE Computer Society.

Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., & Szeliski, R. (2006). A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Computer vision and pattern recognition (CVPR)*. New York, NY: IEEE Computer Society.

Terzopoulos, D. (1988). The computation of visible-surface representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *10*(4), 417–438.

Vlasic, D., Peers, P., Baran, I., Debevec, P., Popovi'c, J., Rusinkiewicz, S., et al. (2009). Dynamic shape capture using multi-view photometric stereo. *ACM Transactions on Graphics (SIGGRAPH-ASIA)*, *28*(5), 1–11.

Wholer, C. (2009). *3D computer vision: efficient methods and applications*. New York: Springer.

Woodham, R. J. (1980). Photometric method for determining surface orientation from multiple images. *Optical Engineering*, *19*(1), 139–144.

Wu, T. P., & Tang, C. K. (2006). *Visible surface reconstruction from normals with discontinuity consideration*. In: Computer Vision and Pattern Recognition (CVPR).

Wu, T. P., Sun, J., Tang, C. K., & Shum, H. Y. (2008). Interactive normal reconstruction from a single image. *ACM Transactions on Graphics (SIGGRAPH-ASIA)*, *27*(5), 1–9.

Xiong, Y., & Shafer, S. (1993). Depth from focusing and defocusing. In *Computer vision and pattern recognition (CVPR)*. New York, NY: IEEE Computer Society.