Shading-Aware Multi-view Stereo

Fabian Langguth[1(✉)], Kalyan Sunkavalli[2], Sunil Hadap[2], and Michael Goesele[1]

[1] TU Darmstadt, Darmstadt, Germany
[2] Adobe Research, San Francisco, USA

Abstract. We present a novel multi-view reconstruction approach that effectively combines stereo and shape-from-shading energies into a single optimization scheme. Our method uses image gradients to transition between stereo-matching (which is more accurate at large gradients) and Lambertian shape-from-shading (which is more robust in flat regions). In addition, we show that our formulation is invariant to spatially varying albedo without explicitly modeling it. We show that the resulting energy function can be optimized efficiently using a smooth surface representation based on bicubic patches, and demonstrate that this algorithm outperforms both previous multi-view stereo algorithms and shading based refinement approaches on a number of datasets.

1 Introduction

High-quality digitization of real world objects has been of great interest in recent years. The demand for effective and accurate digitization methods is increasing constantly to support applications such as 3D printing and visual effects. Passive reconstruction methods such as multi-view stereo [1] are able to achieve high quality results. However, stereo methods typically operate on image patches and/or use surface regularization in order to be robust to noise. As a result, they often cannot recover fine-scale surface details accurately. These details are often captured by shading variations, and recent work has focused on shading-based refinement of the geometry obtained from multi-view stereo (or in some cases using depth sensors or template models). Starting from the work of Wu et al. [2] that can only be used for objects with constant albedo, algorithms have evolved to operate on implicit surfaces [3] and real time settings [4]. All these methods treat the coarse input geometry as a fixed ground truth estimate of the shape and use it to regularize their optimization. Consequently, uncertainties in the inital reconstruction method are discarded and cannot be resolved reliably.

Another challenge for shading-based refinement techniques is that observed image intensities combine shading and surface albedo. Inferring fine-scale detail

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46487-9_29) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part III, LNCS 9907, pp. 469–485, 2016.
DOI: 10.1007/978-3-319-46487-9_29

from shading thus requires reasoning about surface albedo. This significantly increases the number of variables in the optimization. Most current techniques either assume constant albedo or apply strong regularization on the albedo, which can often fail on real-world surfaces.

In contrast to previous work, we propose a new multi-view surface reconstruction approach that combines stereo and shading-based data terms into a single optimization scheme. At the heart of our algorithm is the observation that stereo-matching and shape-from-shading have complementary strengths. While stereo correspondences are more accurate in regions with many large image gradients, shape-from-shading is typically more robust in flat regions with no albedo variations. The resulting algorithm provides three distinct advantages over previous work:

- It leads to a combined multi-view stereo and shading-based reconstruction that balances the two terms without committing, a priori, to either of them.
- It uses a simple image gradient-based trade-off between stereo and shading energies that maximizes their effectiveness.
- It treats spatially varying albedo implicitly, i.e. our optimization is robust against spatially varying albedo without explicitly modeling it.

We show that this combined energy can be optimized efficiently using a continuous surface representation [5]. We demonstrate the effectiveness of this technique on various datasets and show that it outperforms previous MVS and shading-based refinement techniques.

2 Related Work

High-quality surface reconstruction has been an active field of research over the past decade, and approaches have been developed for various forms of input data. Our technique uses an unstructured set of images (with camera parameters) of an approximately Lambertian scene and does not require any special hardware setup. We will review related methods that either operate on similar input data or use ideas similar to our approach.

Multi-view Stereo. Multi-view stereo algorithms [1] are arguably one of the most general passive reconstruction techniques. Approaches such as Goesele et al. [6] and Furukawa and Ponce [7] have shown that geometry can be recovered even for large scale and uncontrolled Internet data. Other approaches use more controlled settings or additional input such as object silhouettes [8]. Multi-view stereo approaches usually add a form of regularization to deal with structureless areas that are not well matched by classical stereo terms such as photo consistency. Similarly regularization is used in two-view stereo methods such as Hirschmueller [9], Bleyer et al. [10], and Galliani et al. [11], which can also be applied to multi-view scenarios by combining many two-view estimates into a robust multi-view estimate. In contrast, our goal is to avoid explicit regularization; instead, we use a new shading-based data term to handle sparsely

textured regions where a traditional stereo term is not very effective. To do this we optimize both depth and normals of a continuous surface. In terms of surface representation, stereo algorithms usually recover a single depth per-pixel [6], a global point cloud [7], or an implicit surface model [8], all of which we found difficult to apply to our approach. Recently another surface representation was proposed inside a multi-view framework by Semerjian [5]. This approach uses bicubic patches to define a surface per view that has continuous depth and normals. We found this representation to be appropriate for our method and adopt it as described later.

Combining Multi-view and Photometric Cues. To recover more detail in regions where depth reconstruction is not very accurate, several methods have combined multi-view and photometric principles. Most of them, however, rely on a controlled and complex capture setup. The approach by Nehab et al. [12] combines two separate reconstructions. They capture depth using structured light, acquire surface normals using photometric stereo, and integrate both these estimates in a separate step. Other approaches such as Hernandez et al. [13] and Zhou et al. [14] combine photometric stereo information from multiple view points into a single framework. This requires a large amount of input data and a complex acquisition system as both light and camera positions need to be controlled. Beeler et al. [15] augment the geometry of captured faces with fine details using the assumption that small concavities in the skin appear darker than flat areas. They do not require a lot of input data but are still dependent on a calibrated capture setup as they do not have a variable lighting model.

Shading-Based Refinement for General Illumination. Most recently, a new line of work uses shading cues from images captured under uncontrolled illumination to improve a given geometry. Wu et al. [2] presented the first approach that uses a precomputed multi-view stereo reconstruction to estimate a spherical harmonics approximation of the lighting. They use this lighting and a shading model to improve the stereo reconstruction. Their approach is able to recover fine-scale details but is limited to objects with a single, constant albedo. Later, Yu et al. [16] and Han et al. [17] both presented algorithms that operate on a single RGB-D input image (e.g., from a Kinect sensor). These sensors usually generate very coarse geometry and shading-based refinement increases the quality and resolution of the output. Xu et al. [18] also extended the idea and developed a simultaneuos opimization of lighting and shape parameters. They do, however, require additional information about the visual hull of the object. Using GPU-based parallel solvers, Wu et al. [19] and Or-El et al. [4] were able to achieve real-time performance on similar input data. All these techniques are still limited to a single albedo [4,17,18], a fixed set of constant albedo clusters [16], or a coarse initial albedo estimate [19]. Other methods focus on more specific scenarios such as faces. Chai et al. [20] fit a parametric face model to an input image and use it for lighting estimation and shading-based refinement. The first technique to include a spatially varying albedo was proposed by Zollhoefer et al. [3]. They include the albedo in the optimization and constrain it using a chromaticity-based regularization scheme similar to Chen and Koltun [21].

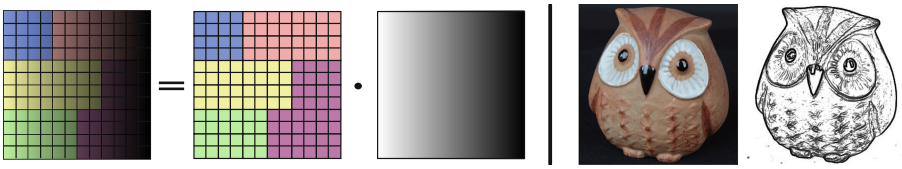

Fig. 1. *Left*: An illustration of our Retinex-based assumption of separating albedo from shading. Large gradients in the image are usually caused by albedo changes; small gradients on the other hand are observed due to lighting. Based on this we compute a trade-off between stereo and shading energies. *Right*: Visualization of the trade-off for an input image. For every pixel we use mainly our stereo term (dark regions) or our shading term (bright regions) based on the magnitude of the image gradient.

While, this prevents shading from being absorbed into the albedo, it can fail in scenes where the albedo variation is not accurately predicted by chromaticities (e.g., albedos with the same chromaticity but different brightness).

Although shading-based refinement techniques have improved significantly in recent years, the basic principle of all existing methods remains the same: They use fixed input geometry, estimate lighting, and later refine the geometry using shading cues. While we also compute a lighting function on a coarse estimate of the geometry, we integrate the geometry refinement directly into the multi-view stereo reconstruction method. This allows us to balance stereo matching and shading cues as we can resolve ambiguities in the multi-view stereo energy, instead of treating the input geometry as fixed. This approach ultimately also enables us to optimize the geometry independent of the (potentially spatially-varying) albedo, i.e., without explicitly including albedo terms into our energy. This is a significant advantage because we do not have to rely on albedo regularization models that can often fail on real-world scenes.

3 Energy Formulation

Our energy balances geometric errors versus shading errors depending on the local image gradient. This is motivated by Land's Retinex theory [22], which assumes that shading introduces only small image gradients, changing the surface brightness gradually. Strong gradients on the other hand are usually caused by changes in surface materials and are thus independent of the illumination. Retinex theory has been commonly used to separate surface albedo and shading [23,24] (see Fig. 1).

In our context, this observation has two implications. First, in multi-view reconstruction the geometric stereo term is usually accurate and robust in regions with strong gradients but fails for small gradients. Many stereo methods therefore use surface regularization to keep textureless areas smooth. We instead utilize the fact that small gradients are most likely caused by lighting and define an additional data term based on a shading function that specifically constrains the direction in which the surface should change. Second, we show that, in regions of

small gradients, we can factor the surface albedo out completely, resulting in an albedo-free shading term. Our error terms are based purely on point wise image gradients and do not involve image values or larger patches of pixels.

The input to our algorithm is an unstructured set of images as well as known camera parameters which can be either pre-calibrated or recovered by stucture from motion tools such as VisualSFM [25]. We aim to compute a depth map for evey view i using a set of neighbor views $j \in \mathcal{N}_i$.

3.1 Geometric Error

Our camera model follows standard definitions [26]. A 3D point $\mathbf{X}$ is transformed into an image location $\mathbf{x}$ in the camera coordinate system according to a camera calibration matrix K, rotation R, and translation $\mathbf{t}$ as

$$\mathbf{x} = \mathrm{K}\left(\mathrm{R}\mathbf{X} + \mathbf{t}\right). \tag{1}$$

For homogeneous coordinates the projection from a pixel coordinate $\mathbf{x}_i$ in camera i into another camera j can then be defined according to a depth value $d_i(\mathbf{x}_i)$ along the principal ray of view i:

$$P_j(\mathbf{x}_i, d_i(\mathbf{x}_i)) = \mathrm{K}_j\left(\mathrm{R}_j\mathrm{R}_i^{-1}\left(\mathrm{K}_i^{-1}\mathbf{x}_i \cdot d_i(\mathbf{x}_i) - \mathbf{t}_i\right) + \mathbf{t}_j\right) \tag{2}$$

The geometric error is now defined as a stereo term based on matching intensity gradients from the main view into neighboring views according to the current depth function. Traditional stereo methods often optimize using image values over a local patch of pixels. Even for illumination invariant measures such as normalized cross-correlation, this would be more difficult to integrate into our Retinex assumption as a patch of pixels is more likely to be affected by both albedo and shading changes. Instead, we specifically optimize this energy for local image gradients. A gradient-based stereo term was introduced by Scharstein [27] but has only been adapted in some specific scenarios like gradient domain rendering [28]. Semerjian [5] recently showed that a point-wise measure of gradients can be very effective for surface reconstruction if used correctly. We adopt this measure as it is well suited for our approach. For any two views i, j and their intensity functions I_i, I_j, and a pixel coordinate $\mathbf{x}_i$ it can be written as:

$$E_g^j(d_i, \mathbf{x}_i) = \nabla I_i(\mathbf{x}) - \nabla I_j(P_j(\mathbf{x}_i, d_i(\mathbf{x}_i))). \tag{3}$$

Here, and in further equations, ∇ denotes image gradients which are the derivatives computed with respect to image coordinates $\mathbf{x}_i$. Note that this also involves the derivative of the projection P_j which transforms the gradient into the correct coordinate system. In addition to constraints beween the main view and its neighbors, we also define pairwise terms between two neighbors as used by Semerjian [5]. Still using the depth of the main view d_i we get:

$$E_g^{j,k}(d_i, \mathbf{x}_i) = E_g^j(d_i, \mathbf{x}_i) - E_g^k(d_i, \mathbf{x}_i) = \nabla I_j(P_j(\mathbf{x}_i, d_i(\mathbf{x}_i)) - \nabla I_k(P_k(\mathbf{x}_i, d_i(\mathbf{x}_i)), \tag{4}$$

where $E_g^{i,j} = E_g^j$. This essentially measures the difference in error between neighbors and avoids overfitting to only one neighbor.

3.2 Shading Error

Lighting Model: Similar to previous work [2,3] we assume Lambertian reflectance. This allows us to define shading as a function of the surface normal $\mathbf{n}$, and independent of the viewing direction. We also use third-order spherical harmonics basis functions B_h to approximate the incoming illumination. The outgoing radiance $R(\mathbf{x})$ at a point $\mathbf{x}$, with albedo $a(\mathbf{x})$ and normal $\mathbf{n}(\mathbf{x})$, is a weighted sum of these bases, which we define as our shading function S:

$$R(\mathbf{x}) = a(\mathbf{x}) \cdot \sum_{h=1}^{16} B_h(\mathbf{n}(\mathbf{x})) \cdot \mathbf{l}_h = a(\mathbf{x}) \cdot S(\mathbf{n}(\mathbf{x}), \mathbf{l}) \tag{5}$$

The lighting parameters $\mathbf{l}$ are computed ahead of surface optimization using a coarse initial surface model derived from basic stereo. This optimization is identical to Zollhoefer et al. [3], i.e., we initialize the albedo as constant and simply solve a linear least squares system. In contrast to Zollhoefer et al., we optimize $\mathbf{l}$ using only our single main image. Using more images would make this estimation more robust, but we explicitly want to optimize for a separate lighting model per image to be invariant to changing light conditions, e.g., an object moving on a turn table or outdoor scenes with uncontrolled lighting. We also set our albedo to a constant value. As we will describe later, we are able to optimize the geometry without explicitly modeling the albedo. This has many advantages for the optimization procedure, but unlike Zollhoefer et al. [3] we cannot create an improved lighting model in further iterations. While there are obvious scenarios that will break this approach, the low number of lighting parameters causes the estimation to be robust enough for a variety of objects, as we will demonstrate in the results. In fact, we observed that in practical scenarios it is much more likely that errors appear due to specular surfaces, self shadowing and inter-reflections, which cannot be dealt with in either case.

Shading Error: Our shading term is also based on image gradients. Similar to [3], we assume that the observed image gradient, ∇I, should be identical to the gradient of the reflected intensity predicted by our model, ∇R, with:

$$\nabla R(\mathbf{x}) = \nabla a(\mathbf{x}) \cdot S(\mathbf{n}(\mathbf{x}), \mathbf{l}) + a(\mathbf{x}) \cdot \nabla S(\mathbf{n}(\mathbf{x}), \mathbf{l}). \tag{6}$$

However, at this point we do not have an accurate model of the albedo. Previous approaches therefore include the albedo in the optimization leading to a significantly bigger, under-constrained problem. This requires an explicit regularization on the albedo using approximate measures such as pairwise differences based on chromaticity. Instead, we use the Retinex assumption to create an albedo independent optimization that does not require any explicit regularization. A common approach for intrinsic images [21,23] is to operate in the log domain as this makes albedo and shading terms additive instead of multiplicative:

$$\log(R(\mathbf{x})) = \log(a(\mathbf{x})) + \log(S(\mathbf{n}(d_i(\mathbf{x})), \mathbf{l})). \tag{7}$$

If we take the gradient with respect to image coordinates we get:

$$\nabla \log(R(\mathbf{x})) = \frac{\nabla a(\mathbf{x})}{a(\mathbf{x})} + \frac{\nabla S(\mathbf{n}(d_i(\mathbf{x})), \mathbf{l})}{S(\mathbf{n}(d_i(\mathbf{x})), \mathbf{l})}. \tag{8}$$

If we now assume—according to the Retinex theory—that small gradients are caused solely by lighting, the albedo gradient vanishes and we can write:

$$\nabla \log(R(\mathbf{x})) = \frac{\nabla S(\mathbf{n}(d_i(\mathbf{x})), \mathbf{l})}{S(\mathbf{n}(d_i(\mathbf{x})), \mathbf{l})}. \tag{9}$$

This means that the difference, $\nabla \log(I(\mathbf{x})) - \nabla \log(R(\mathbf{x}))$, can in fact be minimized by solely optimizing over the shading function, $S(\mathbf{n}(\mathbf{x}), \mathbf{l})$. This indicates an albedo invariance which can also be thought of in the following way: If the albedo is locally constant, an intensity gradient is only caused by a change in surface normals, and given a lighting model, the surface normals have to change in a particular direction which does not depend on the actual value of the albedo. Our shading error is therefore defined as

$$E_s(d_i, \mathbf{x}) = \frac{\nabla I(\mathbf{x})}{I(\mathbf{x})} - \frac{\nabla S(\mathbf{n}(d_i(\mathbf{x})), \mathbf{l})}{S(\mathbf{n}(d_i(\mathbf{x})), \mathbf{l})}. \tag{10}$$

Note that this is a simple point-wise measure which matches the point-wise nature of our gradient-based stereo term and suggests a balanced optimization if both are combined.

3.3 Combined Energy

To formulate our final energy function we combine both data terms in a simple but effective way. For pixels with strong gradients, we rely on the geometric stereo term as it is very robust. For small gradients, we additionally use our shading error as it constrains the surface according to the given lighting model. As we want to do this on a per-pixel basis, we need a continuous trade-off to avoid artifacts. Our solution is to use the magnitude of the image gradient to compute a weight on the shading error term, see Fig. 1 for an example. For a set of neighbors, $\mathcal{N}_i$, including i itself, and a set of pixels, $\mathcal{V}_i$, that are visible in the corresponding neighbors, the final energy is defined as:

$$E(d_i) = \sum_{j,k \in \mathcal{N}_i}^{k>j} \sum_{\mathbf{x}_v \in \mathcal{V}_i} |E_g^{j,k}(d_i, \mathbf{x}_v)| + \frac{\alpha}{\|\nabla I(\mathbf{x}_v)\|_2} |E_s(d_i, \mathbf{x}_v)|, \tag{11}$$

where $\alpha = 0.01$ balances the scale of both terms as the shading error is measured in the log domain. We use the same value for all our datasets. We also experimented with normalizing the weight across pixels. The new weight β would then also affect the geometric error, i.e., $(1-\beta)E_g + \beta E_s$, resulting in a total weight of 1 for each pixel. However, this led to worse results. Note that the final energy is constructed only with local measures and does not contain any explicit regularization terms. Instead it is implicitly regularized by the Retinex assumption

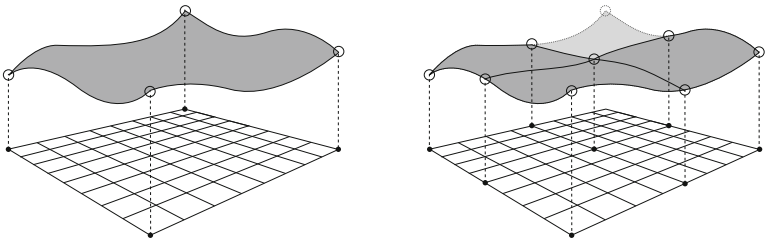

Fig. 2. Surface representation based on bicubic patches. Each patch is defined via 4 nodes (illustrated as circles) that are located at pixel corners (illustrated as dots on the pixel grid). When moving to a higher scale the patch is subdivided and some patches are removed if they have a high error.

and the lighting model. We also use the L1 norm for both our data terms as it is more robust to outliers that do not correspond to our Retinex assumption. It also avoids scale issues in the optimization that can be caused by the shading energy becoming very large in dark areas.

4 Surface Representation and Optimization

As discussed in Sect. 2, we use the framework of Semerjian [5] to optimize our energy function. It provides a surface representation with a continuous definition of depth values and surface normals which is very beneficial for our combined energy. Optimizing a depth map for each view allows us to handle datasets with varying lighting conditions and enables straight forward parallel processing. As this framework uses a different approach compared to simple pixel-wise depth values, we briefly summarize the main aspects.

4.1 Surface Representation

The surface is not represented as depth values per pixel but rather as a set of bicubic surface patches. Every patch is defined by bicubic interpolation between 4 nodes, and neighboring patches share two nodes (see Fig. 2). A node itself represents 4 optimization variables: the depth, the first derivatives of the depth and the mixed second derivative. The nodes are located at image coordinates of the main view and each bicubic patch covers a set of pixels. This also enables an easy formulation of scale, as patches can cover more pixels to represent a coarser scale and can be subdivived to move to a finer scale. At the finest scale the patches cover a 2×2 set of pixels.

4.2 Optimization

Given this representation, we can efficiently optimize the non-linear energy (Eq. 11) using a Gauss-Newton type solver. As our shading error is albedo-free,

we do not need to introduce additional variables and can operate solely on the surface representation. Starting from an initial guess the current energy is linearized, and we solve for an update to the optimization variables. Let $\mathbf{d}$ be the vector of optimization parameters, $\hat{\mathbf{d}}$ the update, and $\mathbf{f}(\mathbf{d})$ the vector of residuals generated by our energy E. Linearizing the error function around the current solution using the Jacobian, $\mathrm{J}_{\mathbf{f}}$, leads to the common linear system:

$$\mathbf{f}(\mathbf{d}+\hat{\mathbf{d}}) \approx \mathbf{f}(\mathbf{d}) + \mathrm{J}_{\mathbf{f}}^{T}\hat{\mathbf{d}}, \qquad \left(\mathrm{J}^{T}\mathrm{J}\right)\hat{\mathbf{d}} = -\mathrm{J}^{T}\mathbf{f} \tag{12}$$

The approximate Hessian $\mathrm{J}^{T}\mathrm{J}$ consists of 4×4 blocks that correspond to the 4 optimization variables at each node. It is also very sparse due to the limited support of the bicubic patches; each node is used for a maximum of 4 patches. The linear system can therefore be solved efficiently using a conjugate gradient solver. The inverse of the block diagonal of $\mathrm{J}^{T}\mathrm{J}$ is a good preconditioner and can be computed quickly using Cholesky decompositions on the blocks.

4.3 Final Algorithm

We first create an initial geometry using the multi-scale formulation and surface operations of Semerjian [5] for coarse scales. Smaller patch sizes of 8×8 and lower are then optimized using our new energy. Applying our shading term for coarse scales would not improve the final result as geometry details are only revealed at finer scales. Another reason is efficiency; the shading error additionally involves the gradient of the shading function and is therefore more complicated to compute which increases the runtime compared to simple regularization. Finally, the reconstructed surfaces from all views are converted to a point set with normals and can be fused with any surface reconstruction algorithm [29–31]. Each view can also be represented as a depth or normal map.

5 Results

In the following, we evaluate our method using a variety of datasets. For all our results we used 6–9 neighbor images (except for the sparse Middlebury datasets) and fused them into a global model using Floating Scale Surface Reconstruction (FSSR) [30]. We chose this approach because it does not fill holes that may appear in the geometry due to large errors in our stereo and/or shading energy.

We first evaluate our approach on the well known Middlebury benchmark [1]. Comprehensive results are available on the website. The *Dino* dataset has many areas that are affected by self shadowing and interreflections. As Fig. 3 shows, our optimization can handle these effects in many cases if enough stereo information from multiple views is available. Note that our optimization handles cast shadows to some extent implicitly since the weight for the shading term is low at the shadow boundaries, and cast shadows can be matched well with stereo matching. The lighting model is, however, still wrong inside the shadowed

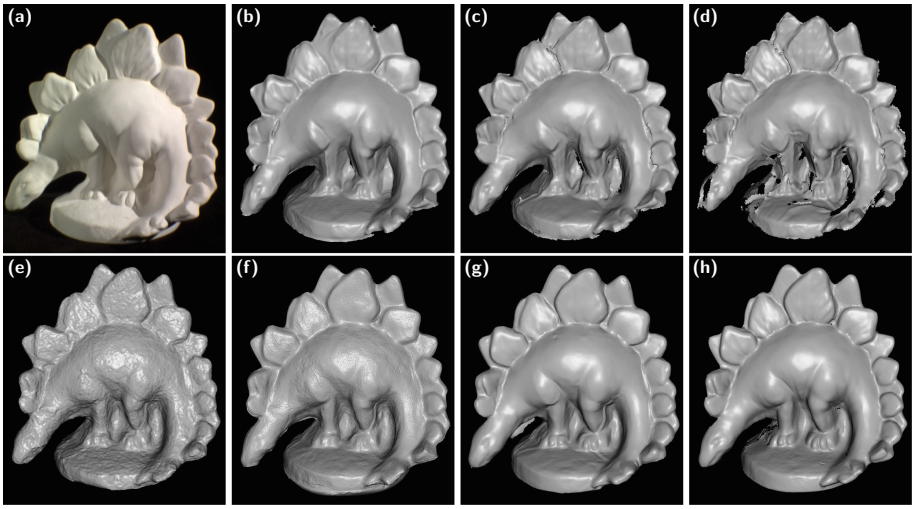

Fig. 3. Results on the *Dino* dataset of the Middlebury benchmark with decreasing number of input images. This dataset has strong shadowing which can be seen in the input image. However, in areas where our lighting model is correct we are able to recover a high amount of detail in the geometry even for sparse input data. *Top*: (a) Input image; (b) our reconstruction on full dataset, 363 images, using 9 neighbors; (c) ring dataset, 46 images, using 4 neighbors; and (d) sparse ring dataset, 16 images, using only 1 or 2 neighbors. *Bottom*: Results on full dataset submitted by (e) Furukawa et al. [7], (f) Galliani et al. [11], and (g) Semerjian [5]; and (h) ground truth.

areas since the incoming illumination is partially occluded. On the full dataset our result has an accuracy of 0.49 mm and a completeness of 96.9 %. For the sparse *Dino* dataset where stereo cues are not very strong, our shading term causes holes in the shadowed areas as we cannot find consistent normals in these areas. However, compared to other approaches, we are able to recover a significant amount of detail in areas that are not affected by shadows. In fact, we reconstruct the same amount of detail independent of the sparsity of the input data, which highlights another strength of our shading term. Even for the very sparse input data of 16 images and using only 2 neighbors we can reconstruct more detail than top scoring approaches on the full dataset. For the full *Temple* dataset (Fig. 4), we are able to achieve a high accuracy even though the back of the object has many concavities leading to strong interreflections that cannot be represented by our global lighting model. Compared to the results submitted by Semerjian [5] our shading term improves the accuracy on the full dataset by 0.15 mm to 0.47 mm and we achive a completeness of 98.7 %.

Figures 5 and 6 show *fountain-P11*, an outdoor dataset from the Strecha et al. [32] benchmark. The normal maps in Fig. 5 show the effect of different surface regularization weights on the original approach of Semerjian [5]. There is no globally correct weight as the reconstructed geometry is either too smooth or

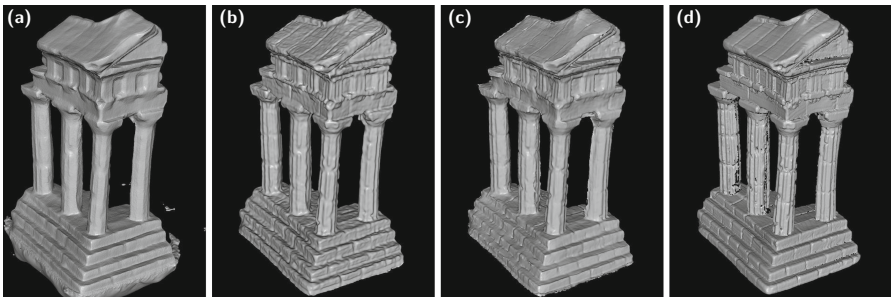

Fig. 4. Results on the *Temple* dataset of the Middlebury benchmark. *From left to right*: (a) Galliani et al. [11]; (b) Fuhrmann et al. [30] using the stereo from Goesele et al. [6]; (c) our reconstruction; and (d) ground truth. Our reconstruction achieves a good balance between capturing fine-scale detail without introducing noise.

too noisy. In contrast, our approach reconstructs smooth but detailed geometry due to the image gradient magnitude-based weight. Figure 6 demonstrates that this also translates to the fused geometry as integrating multiple views cannot remove the noise inherent in Semerjian's reconstruction effectively.

Next we present a multi-scale outdoor dataset included in the FSSR paper [30]. Figure 7 shows that our approach can recover detailed geometry in such a setting. The normal map captures even the finest details recovered in a single view. Our results from vastly different scales can be combined into a consistent model with FSSR. However, we can observe the boundaries between scales as the resolution and accuracy of the geometry changes drastically. This still illustrates an advantage compared to other systems that operate on a global model: our approach can scale to any amount of images and can easily reconstruct different levels of detail in a single dataset, whereas keeping a multi-scale global model in an efficient data structure is challenging and not arbitrarily scalable.

Figure 8 shows a dataset presented by Zollhoefer et al. [3]. This object already provides many gradients for stereo matching so we do not expect our shading term to result in a substantial improvement. Note, however, that our reconstruction has significantly better quality compared to the normal map reconstructed with Semerjian's approach, and compared to the Zollhoefer et al. [3] reconstruction provided on their project web page.

Finally, Fig. 9 presents results on a dataset captured under varying lighting conditions. The *Owl* was captured on a turn-table with fixed lights and a fixed camera, resulting in different lighting for each image (w.r.t. the image coordinates). The object is nearly diffuse apart from the dark specular areas where all the methods shown here fail. We compare against a patch based stereo method [6], which has no effective regularization as each pixel is optimized independently. This results in a very uneven surface and noise in (almost) textureless regions. Semerjian [5] uses a simple regularization term that keeps the surface variation low. This is effective in producing a continuous surface, but cannot recover

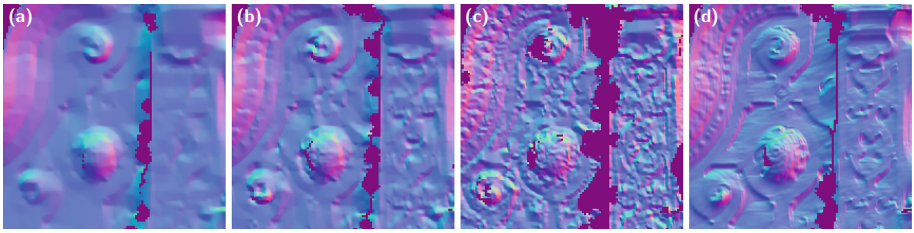

Fig. 5. The *fountain-P11* dataset from Strecha et al. [32]. *From left to right*: Closeup normal maps for single views of the bottom left area for different weights on surface regularization (a) high, (b) medium, and (c) low; and (d) normal map of our reconstruction. Basic regularization cannot find a good trade-off between overly smooth and noisy geometry. Our result reveals fine details without introducing noise.

Fig. 6. The *fountain-P11* dataset from Strecha et al. [32]. *From left to right*: Reconstruction by our implementation of Semerjian [5] using a low regularization weight to recover details; by our new optimization; and ground truth.

details in regions without strong gradients. In contrast, our combined method recovers a smooth surface and is able to relate small gradients to surface details.

5.1 Runtime

A C++ implementation of our technique is available as open source software[1]. This unoptimized prototype shows a roughly 20 % runtime increase compared to our implementation of Semerjian [5]. In practice, the full *Dino* and *Temple* datasets were computed in 75 and 63 min on a 32-core machine. The multi-scale outdoor dataset from Fuhrmann et al. [30] included 204 high resolution images and was computed in 115 min on the same machine, while the *Owl* dataset with 10 images took around 7 min. For a fair comparison to other stereo methods, we are reporting the run-times of our complete multi-view algorithm and not only the time required for solving our shading-based optimization.

[1] https://github.com/flanggut/smvs.

Fig. 7. Results on an outdoor dataset. *Top*: Input images at different scales, and our global model with details. Our method recovers more detail in regions that are imaged at higher-resolution. *Bottom*: A closeup input image; the reconstructed depth map shaded with the lighting; and close-up normals with regular (10^{-2}) and low (10^{-4}) value for α – decreasing the weight of the shading term results in more noise and less detail as the stereo term dominates the energy.

5.2 Limitations

We make two main assumptions in our method that can lead to errors in the final geometry if they are violated. First, we assume that the scene is Lambertian and a low frequency spherical harmonics lighting can accurately represent the illumination. As we show in the Middlebury *Dino* dataset, shadows and interreflections will cause errors in the reconstruction but we are still able to reconstruct details in areas where our lighting model is correct. A more sophisticated lighting model could solve the issues in future work, and would require only minor changes to our geometry optimization. Second, we assume that we can separate albedo and lighting according to the magnitude of the image gradient. While this holds for many datasets, there are objects where the albedo changes gradually, and this violation of Retinex can show up in our geometry if we relate these small gradients to shading and therefore changes in the surface normal. This suggests that some geometry regularization might still be needed in certain regions where we cannot easily decide between albedo and shading. As we rely solely on the stereo error for strong gradients, we are also limited by its accuracy. In certain configurations, e.g., observing horizontal lines under horizontal camera motion, or

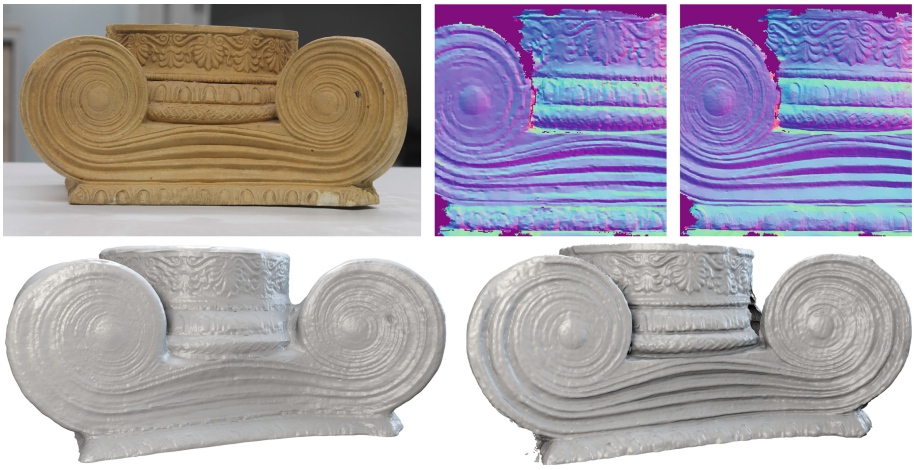

Fig. 8. The *Figure* dataset. *Top from left to right*: An input image of the dataset; normal maps from the surface computed by our implementation of Semerjian [5], and our shading based approach. *Bottom*: Result presented by Zollhoefer et al. [3] (available at project website); and our fused model.

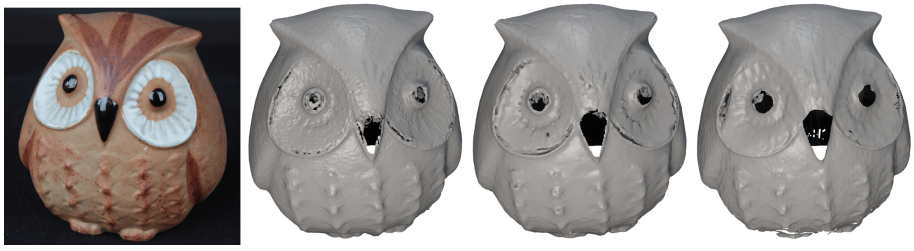

Fig. 9. Reconstruction of the *Owl* dataset with changing lighting in each image. *From left to right*: An input image; reconstruction by Goesele et al. [6]; by our implementation of Semerjian [5]; and using our new optimization. Our results capture more structural details (see the eyes for example) with less overall noise.

fine structures with aliasing effects, the stereo term might lead to wrong depth estimates that we cannot fix with our normal-based shading term.

6 Conclusion

In this paper, we have presented a novel multi-view surface optimization algorithm that efficiently combines a stereo energy term with a shading-based energy term in a single, combined approach, to create high quality reconstructions. Building on the Retinex assumption, we are able to completely remove the albedo from the shading-based error, which has not been done before. Our formulation relies solely on pixel-wise data terms and an implicit regularization

via the shading function and surface representation. We present results that improve on previous multi-view stereo algorithms and shading based refinement systems. Our approach is limited by the basic lighting model and cannot account for self-shadowing, indirect illumination, and specular materials. In future work we will improve on this to create a more robust system that can be applied to more complex scenes. Overall, we believe that the idea of combining stereo and shading energies can be very powerful and will lead to more general approaches.

Acknowledgements. This work was supported in part by the European Commissions Seventh Framework Programme under grant agreements no. ICT-611089 (CR-PLAY).

References

1. Seitz, S., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), pp. 519–526. IEEE Computer Society (2006)
2. Wu, C., Wilburn, B., Matsushita, Y., Theobalt, C.: High-quality shape from multi-view stereo and shading under general illumination. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011 (2011)
3. Zollhöfer, M., Dai, A., Innmann, M., Wu, C., Stamminger, M., Theobalt, C., Nießner, M.: Shading-based refinement on volumetric signed distance functions. ACM Trans. Graph. **34**(4), 96:1–96:14 (2015). doi:10.1145/2766887. article no 96
4. Or-El, R., Rosman, G., Wetzler, A., Kimmel, R., Bruckstein, A.M.: RGBD-fusion: real-time high precision depth recovery. In: Computer Vision and Pattern Recognition (CVPR) (2015)
5. Semerjian, B.: A new variational framework for multiview surface reconstruction. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 719–734. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10599-4_46
6. Goesele, M., Snavely, N., Curless, B., Hoppe, H., Seitz, S.: Multi-view stereo for community photo collections. In: International Conference on Computer Vision (ICCV) (2007)
7. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multi-view stereopsis. Trans. Pattern Anal. Mach. Intell. (PAMI) **32**(8), 1362–1376 (2010)
8. Heise, P., Jensen, B., Klose, S., Knoll, A.: Variational patchmatch multiview reconstruction and refinement. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 882–890 (2015)
9. Hirschmüller, H.: Accurate and efficient stereo processing by semi-global matching and mutual information. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2005, Washington, DC, USA, pp. 807–814. IEEE Computer Society (2005)
10. Bleyer, M., Rhemann, C., Rother, C.: Patchmatch stereo - stereo matching with slanted support windows. In: Proceedings of the British Machine Vision Conference, pp. 14.1-14.11 (2011)
11. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 873–881 (2015)

12. Nehab, D., Rusinkiewicz, S., Davis, J., Ramamoorthi, R.: Efficiently combining positions and normals for precise 3D geometry. In: ACM SIGGRAPH 2005 Papers (2005)
13. Hernandez Esteban, C., Vogiatzis, G., Cipolla, R.: Multiview photometric stereo. IEEE Trans. Pattern Anal. Mach. Intell. **30**(3), 548–554 (2008)
14. Zhou, Z., Wu, Z., Tan, P.: Multi-view photometric stereo with spatially varying isotropic materials. In: Computer Vision and Pattern Recognition (CVPR), pp. 1482–1489 (2013)
15. Beeler, T., Bickel, B., Beardsley, P., Sumner, B., Gross, M.: High-quality single-shot capture of facial geometry. ACM Trans. Graph. **29**(4), 40:1–40:9 (2010). doi:10.1145/1778765.1778777. article no 40
16. Yu, L.F., Yeung, S.K., Tai, Y.W., Lin, S.: Shading-based shape refinement of RGB-D images. In: Computer Vision and Pattern Recognition (CVPR) (2013)
17. Han, Y., Lee, J.Y., Kweon, I.S.: High quality shape from a single RGB-D image under uncalibrated natural illumination. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 1617–1624 (2013)
18. Xu, D., Duan, Q., Zheng, J., Zhang, J., Cai, J., Cham, T.J.: Recovering surface details under general unknown illumination using shading and coarse multi-view stereo. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1526–1533 (2014)
19. Wu, C., Zollhöfer, M., Nießner, M., Stamminger, M., Izadi, S., Theobalt, C.: Real-time shading-based refinement for consumer depth cameras. ACM Trans. Graph. **33**(6), 200:1–200:10 (2014)
20. Chai, M., Luo, L., Sunkavalli, K., Carr, N., Hadap, S., Zhou, K.: High-quality hair modeling from a single portrait photo. ACM Trans. Graph. **34**(6), 204:1–204:10 (2015)
21. Chen, Q., Koltun, V.: A simple model for intrinsic image decomposition with depth cues. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 241–248 (2013)
22. Land, E.H.: The retinex theory of color vision. Sci. Am. **237**(6), 108–128 (1977)
23. Horn, B.: Determining lightness from an image. Comput. Graph. Image Process. **3**(1), 277–299 (1974)
24. Grosse, R., Johnson, M.K., Adelson, E.H., Freeman, W.T.: Ground truth dataset and baseline evaluations for intrinsic image algorithms. In: Computer Vision, pp. 2335–2342. IEEE (2009)
25. Wu, C., Agarwal, S., Curless, B., Seitz, S.M.: Multicore bundle adjustment. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3057–3064 (2011)
26. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2004). ISBN: 0521540518
27. Scharstein, D.: Matching images by comparing their gradient fields. In: Proceedings of the 12th ICPR International Conference on Pattern Recognition, vol. 1, pp. 572–575 (1994)
28. Kopf, J., Langguth, F., Scharstein, D., Szeliski, R., Goesele, M.: Image-based rendering in the gradient domain. ACM Trans. Graph. **32**(6), 199:1–199:9 (2013). doi:10.1145/2508363.2508369. article no 199
29. Kazhdan, M., Hoppe, H.: Screened poisson surface reconstruction. ACM Trans. Graph. **32**(3), 29:1–29:13 (2013). doi:10.1145/2487228.2487237. article no 29
30. Fuhrmann, S., Goesele, M.: Floating scale surface reconstruction. In: Proceedings of ACM SIGGRAPH (2014)

31. Ummenhofer, B., Brox, T.: Global, dense multiscale reconstruction for a billion points. In: IEEE International Conference on Computer Vision (ICCV), December 2015
32. Strecha, C., von Hansen, W., Gool, L.V., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8 (2008)