Multi-View Photometric Stereo by Example

Jens Ackermann[1], Fabian Langguth[1], Simon Fuhrmann[1], Arjan Kuijper[2], Michael Goesele[1]
[1]TU Darmstadt [2] Fraunhofer IGD

Abstract

We present a novel multi-view photometric stereo technique that recovers the surface of textureless objects with unknown BRDF and lighting. The camera and light positions are allowed to vary freely and change in each image. We exploit orientation consistency between the target and an example object to develop a consistency measure. Motivated by the fact that normals can be recovered more reliably than depth, we represent our surface as both a depth map and a normal map. These maps are jointly optimized and allow us to formulate constraints on depth that take surface orientation into account. Our technique does not require the visual hull or stereo reconstructions for bootstrapping and solely exploits image intensities without the need for radiometric camera calibration. We present results on real objects with varying degree of specularity and show that these can be used to create globally consistent models from multiple views.

1. Introduction

Image-based reconstruction is a well-researched area of computer vision. Significant progress has recently been made to extend (multi-view) stereo and photometric stereo methods to more general settings. Our goal is to recover the surface of objects with non-Lambertian BRDFs. Reconstructing accurate geometry for such objects is still a very challenging task under unknown lighting conditions if no special setups such as ring lights or calibration steps are employed. For textured objects, techniques such as (multi-view) stereo achieve reconstructions of good quality. Instead, we focus on challenging textureless objects where photoconsistency tests such as NCC or SSD fail. Classical photometric stereo, in contrast, works well in textureless regions but cannot directly recover depth information.

To address these issues, we place a reference object (the "example") with known geometry in the scene. This makes a calibration of the camera response unnecessary which is required by many photometric stereo techniques. We match per-pixel appearance profiles from varying viewpoints with different illumination, using the matching error between the example and target object as consistency measure. While this error is not very discriminative for reconstruction of depth, we show that normals can be recovered very accurately in the vicinity of the true surface.

This approach eliminates several restrictions of the voxel coloring-based work by Treuille *et al*. [29]. Most notably, we operate with general camera and light source positions and use reliably recovered normals as soft constraints for depth recovery. We also reconstruct per-view depth maps instead of a voxelized global model, which has several advantages: There is no need to choose the size of the voxel grid as we work with natural pixel resolution. This leads to less memory consumption, and the algorithm is trivially parallelizable over the individual views. The resulting depth maps can be integrated using standard mesh-merging techniques. In contrast to other multi-view photometric stereo approaches [18, 33, 24, 32, 8, 14], we do not need to separately estimate an intermediate proxy geometry (using other approaches) from which the true surface has to be obtained later on in an additional refinement step. Instead, we couple geometry and normal reconstruction and recover a surface directly from the input data. Our contributions are:

- We present a novel multi-view photometric stereo technique based on matching per-pixel appearance profiles, which makes no assumption about the placement of distant light sources or cameras.
- We analyze the relation between matching ambiguity and normal errors in the multi-view setting and develop an energy formulation that exploits the fact that normals can be recovered more reliably than depth.
- Our technique uses an example object to handle arbitrary uniform BRDFs and also avoids any light or radiometric camera calibration. It thus removes the common assumption of a linear camera response which is often hard to obtain accurately.

We proceed by discussing previous works in this area. We then motivate and explain our approach in Section 3 and provide implementation details in Section 4. Finally, we evaluate our results in Section 5 and close with a conclusion.

2. Related Work

Photometric Stereo: Research related to photometric stereo has started in the eighties with the initial work by Woodham [30]. It relies on varying image intensities to estimate surface orientation and has since then been generalized in many ways. One main direction of research is concerned with jointly recovering unknown shape and reflectances [6, 2, 11, 27]. Another direction focuses on a less restrained capture setup with arbitrary and unknown illumination [7, 26, 23]. Only few works address both challenges simultaneously. They often rely on pixel intensity profiles, as we do. An elegant solution was proposed by Silver [28] and popularized by Hertzmann and Seitz [9, 10]. They place a reference object in the scene and match profiles with the target. We draw inspiration from these works, which make light calibration unnecessary and can handle arbitrary reflectance properties. Similar approaches that do not require a reference object have been presented by Sato *et al.* [25] and Lu *et al.* [19]. They exploit the geodesic distance of intensity profiles and its relation to surface shape.

Single Image Reconstructions: Shape from shading and intrinsic image decomposition methods, *e.g.* [21, 3], operate on single images. They require stronger regularization to compensate for less available data. Johnson and Adelson [13] calibrate against a sphere with the same BRDF similar to our setup. Like the other shape from shading techniques, it could be applied to each view individually in a multi-view setting. Such an approach would, however, be unable to exploit parallax for depth estimation. Extensions to multiple images usually require the depth to be known beforehand (*e.g.* Laffont *et al.*'s intrinsic image technique [17]) and/or a fixed, calibrated lighting environment as presented by Oxholm and Nishino [22].

Multi-View Photometric Reconstructions: Approaches that fuse multi-view cues with photometric stereo are faced with the challenge of finding correspondences between pixels in different images. However, if these were known accurately the problem of shape reconstruction would already be solved. Therefore, most techniques rely on some kind of proxy geometry that gets refined using shading information. Lim *et al.* [18] use a piecewise-planar initialization constructed from tracked feature points. Other common choices are depth maps from structured light [33], multiview stereo reconstructions [24], simple primitive meshes [32], and the visual hull computed from silhouettes [8]. None of these approaches use photometric cues for depth estimation. Furthermore, feature extraction, *e.g.* [31], or stereo reconstruction, *e.g.* [5], fail for textureless objects. The visual hull only provides an adequate initialization if the object is observed from considerably varying angles.

Jin *et al.* [12] use a rank constraint on the radiances in a surface patch collected over multiple images to estimate depth. They assume constant illumination in all images whereas photometric stereo methods exploit the variation of the lighting. Only few works attempt to use varying photometric information for depth estimation. Recently Zhou *et al.* [34] have presented an appearance acquisition method that collects iso-depth contours obtained by exploiting reflectance symmetries in single views. This requires multiple images from the same viewpoint and a calibrated lighting setup. In our case, the camera and light can both move freely. Joshi and Kriegman [14] use the rank-3 approximation error as an indicator of surface depth but are limited to diffuse surfaces. A graphcut optimization is then applied to obtain a discrete depth map as initialization for photometric stereo. Finally, both sources are fused using the integration scheme presented by Nehab *et al.* [20]. In contrast, we do not need the reflectance to be represented as a rank-3 matrix and our surface optimization is directly coupled with the actual image information: We use intensities even during integration similar to Du *et al.* [4] who define a combined energy in a two-view setting. An important difference that sets us apart from all other works that do not rely on intensity profile matching is that any kind of radiometric calibration or linear image intensities becomes unnecessary.

Only one other work approaches the multi-view photometric stereo problem by exploiting an example object: Treuille *et al.* [29] employ the error of matching appearance profiles as introduced by Hertzmann and Seitz [9] and use it as consistency measure in a voxel coloring framework. This approach has, however, several drawbacks: First, it poses restrictions on camera placement to ensure that occluded voxels are processed in the correct order. We allow arbitrary (distant) camera placements and rely solely on generic outlier removal to handle occlusions and shadows. Second, their final scene representation is a voxel grid. The reconstruction cannot be transformed into a surface and the normals can only be used for rendering. Most importantly, their approach cannot use the more reliable normal information during depth recovery, which makes it prone to errors in the reconstructed geometry. Our approach differs from [29] in scene representation (voxels vs multiple depth maps), visibility handling (geometric vs outlier-based), and the reconstruction algorithm (voxel coloring vs per-view optimization).

3. Approach

Our goal is to recover the surface of a textureless object solely from a set of images under varying illumination and from different viewpoints. We also want to keep the capture procedure simple and straightforward. In practice this means to avoid any calibration of light sources or camera response curves. If we also allow for non-diffuse surfaces, none of the existing techniques can be applied. We base our approach on orientation consistency as a depth cue which brings many of the desired properties and thus place a reference object with known geometry in the scene (Figure 1).

Let $I \in \{I_1, \ldots, I_m\}$ denote a master image and r the ray corresponding to pixel p. We assume that the camera projection operators $\{P_1, \ldots, P_m\}$ are known. For a depth

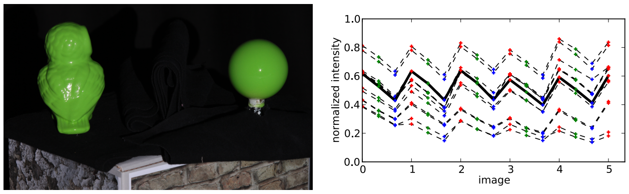


Figure 1. *Left*: Target object and a reference sphere with same reflectance. The high-frequency pattern at the bottom is used to estimate camera pose. *Right*: Some samples from the database of reference profiles (dashed) and a candidate profile (solid).

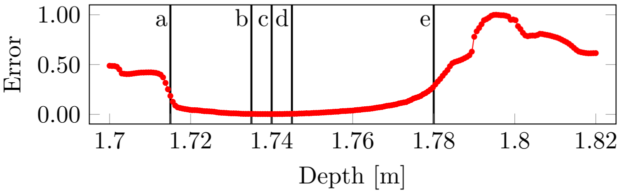


Figure 2. The error of best matching reference profiles along a ray from the camera has a wide basin with very similar error scores. The vertical lines correspond to the depth values in Figure 5.

candidate d we project its 3D position $d \cdot r$ into all m images to obtain intensities $I_j(P_j(dr)), j \in \{1, \dots, m\}$ in each of the three color channels. We call the concatenation of the $3m$ values into a vector $A(dr)$ an *appearance profile*.

As a reference object we use a sphere with known position and radius. In theory, it should have the same reflectance properties as the target object but Section 5 shows that this assumption can be relaxed in practice. For each pixel in I that is covered by the sphere, we project the corresponding sphere point into all images and form a reference appearance profile B. This yields a database of profiles with attached normals $\tilde{n}$ computed from the sphere. Some of these reference profiles B together with a candidate profile A are visualized in Figure 1.

We assume a distant but otherwise unknown point light source L_j. Shadows and inter-reflections are handled as outliers during matching without explicit treatment.

3.1. Appearance Matching

Assuming an orthographic camera, the intensity of a surface point dr with normal n is given by

$$I_j(P_j(dr)) = f_j\left(\int L_j(\omega)\rho(\omega, v_j, n)\langle n,\ \omega\rangle d\omega\right) \quad (1)$$

with camera response f_j, BRDF ρ, and camera viewing direction v_j. Both, light and camera position, change from image to image as indicated by the index j. Note that the right hand side depends only on the normal and not the 3D position. Thus, for a point with the same normal on the surface of the reference object the intensity is the same. This observation is called *orientation consistency*.

This means that we can find a matching profile B in our database for any $A(dr)$ that originates from the true surface. For a false depth candidate d it is unlikely to find a good match, because each view actually observes a different point on the surface. We denote the intensity residuals $e_j = A_j - B_j$ and omit the color channel indexing for simplicity.

Treuille *et al.* [29] use the normalized L_2 distance as a matching error. The contribution of e_j is not considered during matching if the corresponding voxel would actually be occluded in image I_j. We do not have occlusion information available for the components of the target profiles A. Instead, we turn off residuals e_j if the corresponding normal to the reference B would have been observed at a grazing angle in the j-th view. Furthermore, we only use the K best of the remaining residuals:

$$E_{\text{match}}(A, B) = \frac{1}{K}\sum_{i}^{K} e_{j_i}^2. \quad (2)$$

K is a percentage of all views, typically 60%, which acts as outlier handling. For $K < 3$, we set $E_{\text{match}}(A, B) = \infty$, because normals cannot be recovered unambiguously.

3.2. Energy Formulation

Along a ray r the best matching error at position dr

$$E_M(r, d) = \min_B E_{\text{match}}(A(dr), B) \quad (3)$$

gives an indication whether we are on the true surface or not. Unfortunately, the matching error is not very discriminative as shown in Figure 2. We do not observe a clear minimum but rather depth values with a wide basin of low error. Accordingly, choosing the depth with smallest matching error leads to a very inaccurate and noisy depth map. The standard way to deal with noise and unreliable estimates, *e.g.* in stereo, is to employ regularization that favors smooth surfaces. We have the advantage of additional information in the form of normals associated with the best match from the database. To exploit these, we formulate an energy that is defined on both a depth map D and a normal map N. This can be interpreted as attaching a small oriented plane $(D(p), N(p))$ to each ray, see Figure 3, and allows us to encourage integrability without strictly enforcing it since this would be harmful at depth discontinuities.

The key finding in our setting is that exactly the same reasons that make depth estimation hard make normal estimation easy. Figure 4 illustrates this insight for three different points along the same ray. In Figure 4a all cameras observe the same point on the true surface. The matching error will be low and the normal $\tilde{n}$ associated to the match is the correct surface orientation n. If we move slightly away from the surface as shown in Figure 4b, each camera actually observes a different surface point but with normals that are still close to the true one. Accordingly, the intensity profile will be very similar to the previous one. Thus, the matching error is again low which makes accurate depth

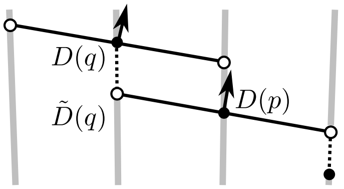


Figure 3. Each ray has a little plane attached. The estimated depth of neighboring pixels should be close to the intersections of their rays with the plane.

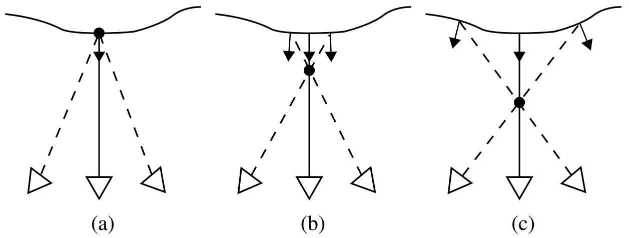


Figure 4. Projections at different depth. (a) All cameras observe the same point. The matching error is zero. (b) Cameras observe different points, but with similar normals. The matching error is still low. (c) Cameras observe points with significantly different normals. The matching error is high.

estimation so difficult, but the associated normal is close to n. This reasoning breaks down if the point is really far away from the surface as in Figure 4c. All cameras observe surface points with very different normals and the normal associated with the best match will not be close to any of them. In this case the matching error itself is high.

Figure 5 shows this effect on real data. For a ray indicated by the dot at pixel p, the best matching normals are visualized for 5 depth values corresponding to the plot in Figure 2. We observe that the normals are almost constant in the region of low error. To exploit this finding we focus our optimization on the normals and use the matching error only as a weak constraint. Based on these considerations, we propose the following energy formulation

$$E(D,N) = E_M(D) + \alpha E_{\text{copy}}(D,N) + \beta E_{\text{coupling}}(D,N). \tag{4}$$

$E_M(D)$ is the sum of matching errors over all rays for the current depth estimates, which involves matching against the intensity database for a single evaluation of $E_M(r,d)$:

$$E_M(D) = \sum_r E_M(r,d)^2. \tag{5}$$

The second term effectively copies the normal $\tilde{n}$ associated to the best matching reference profile, *i.e.* $B = \arg\min E_{\text{match}}$, to the current estimate $n = N(r(p))$ in the normal map but also allows for deviations from the discretely sampled normals on the sphere:

$$E_{\text{copy}}(D,N) = \sum_r \|n - \tilde{n}\|^2. \tag{6}$$

The best matching $\tilde{n}$ also depends on the depth d which we omitted here for clarity. Internally, we parametrize the normals in angular coordinates to ensure unit norm.

The third term couples depth and normals. We assume that the surface is locally planar at a pixel p, but not necessarily fronto-parallel. Since real cameras only approximate an orthographic projection, we consider perspective rays here that all originate at the camera center. We look at a neighboring pixel $q \in \mathcal{N}(p)$ and intersect its ray $r(q)$ with the plane defined by $(D(p), N(p))$

$$\tilde{D}(q) = D(p)\frac{\langle r(p),\ N(p)\rangle}{\langle r(q),\ N(p)\rangle} =: D(p)\frac{s(p)}{s(q)}. \tag{7}$$

The intersection point $\tilde{D}(q)r(q)$ should then be close to the current estimate $D(q)r(q)$ as shown in Figure 3. After multiplication with the denominator we obtain the following coupling term

$$E_{\text{coupling}}(D,N) = \sum_p \sum_{q\in\mathcal{N}(p)} E_{\text{coupling}}(p,q), \tag{8}$$

$$E_{\text{coupling}}(p,q) = \Big(D(p)s(p) - \tilde{D}(q)s(q)\Big)^2. \tag{9}$$

The energy completely and only depends on the actual captured image intensities. This is in contrast to approaches that start with a proxy geometry and then obtain the final surface through a refinement step [14, 24]. Those exploit the additional knowledge about the surface orientation only in this final phase after fundamental decisions on depth have already been made. This can lead to problems if the initialization is inaccurate as in our case. Therefore, we make all decisions at the same time and relate depth and normals directly to the input intensities.

4. Implementation and Experiments

Optimization: We use the Ceres [1] non-linear optimization package to minimize the energy in Equation (4) with the Levenberg-Marquardt algorithm. However, our formulation is non-convex and has many local optima. It is therefore crucial to obtain a sufficiently good initialization for the optimization. We define a depth range which we sample in discrete steps similar to a plane sweep and evaluate only the term E_M. For each pixel we use the depth that results in the lowest error and copy the corresponding normal from the reference object. As already mentioned, these estimates are rather noisy in depth. Still, the normals provide a suitable starting condition. Furthermore, we allow the solver to make jumps that temporarily increase the energy if it ultimately leads to a smaller error. This helps to avoid local optima at the cost of increased run time. We found a total iteration count of 50 to be a good trade-off between quality and computation time. This already decreases the energy by one to two orders of magnitude, *c.f.* Table 1, and we did not

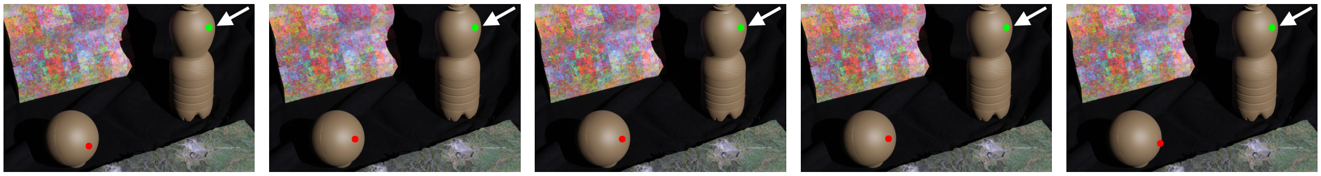


Figure 5. Along the ray going from the camera through the pixel marked in green, the normal corresponding to the best matching reference profile is visualized (red) for increasing depth. Images from left to right correspond to depth a-e in Figure 2. Close to the surface, normals are very stable and similar to the true one.

Dataset	Pixels	Energy after iteration			Time
	in mask	0	10	50	[min]
Bottle	29k	3263	1129	164	459
Diffuse Owl	48k	7712	2408	562	286
Shiny Owl	13k	12331	274	46	130
Spheres	12k	589	49	47	41

Table 1. Computation times and optimization performance.

observe significant improvements through more iterations. Figure 10 illustrates the initialization and the final result. In our prototype, we use images of size 1400×930 and 700×465. This is to reduce run time since the main bottleneck lies in the matching of each candidate profile against all reference profiles. Acceleration with spatial data structures is difficult, because our matching is not a true metric due to the outlier tolerance.

Assumptions in Practice: In Section 3 we made the assumptions that camera parameters and the position of the reference sphere are known. To obtain these parameters, we place a target with a high frequency texture in the scene, see Figure 5. We then extract features and apply structure from motion followed by bundle adjustment. The reference sphere is located by fitting conics to the outline of the sphere in the images. Afterwards, the rays through the sphere center are intersected to find its position. This procedure has the additional advantage of providing us with metric scaling information based on the known radius of the sphere. The metric coordinate system then helps to define the depth range during initialization of the optimization.

Preprocessing: Including all possible images in the reconstruction of a given master view not only leads to increased processing cost, but it can also reduce robustness. If the parallax between two views is too large, chances are that they actually observe different parts of the surface. We avoid measuring consistency between such views and automatically discard images with a viewing direction that deviates more than 50° from the master view. In addition, we manually define a mask for the object in the master view.

Parameter Settings: The weighting factors in Equation (4) are chosen according to the range of each sub-term. The input intensities and E_M are in $[0, 1]$. E_{copy} is in $[0, 2]$ since we do not enforce front-facing normals. We assume that depth is measured in meters, but the typical deviations between neighboring pixels are only fractions of millimeters. Therefore, we scale E_{reg} to lie in a similar range as E_M and E_{copy}. In summary, we set $\alpha = 1$ and $\beta = 5000$ in all our experiments. For much larger β the surface moves away from its true position whereas much smaller values result in more noise. Another parameter is the depth range for the initialization. We manually select a range that encloses the object by 10-15 cm and sample it in 200 steps.

5. Results

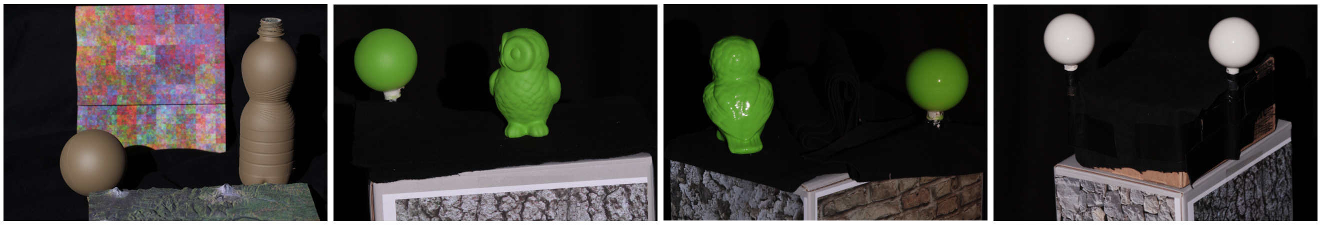
5.1. Experimental Setup

For all experiments we used a point light source at a distance of 5 m to approximate distant illumination. We placed the reference and target objects close together to ensure equal lighting conditions. Figure 6 shows some examples of the input images. The *bottle*, *shiny owl*, and *spheres* datasets were captured by moving the camera and light source in each shot and contain ~15 images. For the *diffuse owl* dataset we captured 39 views from 360° using a turntable. We used a Canon EOS 5D except for the *bottle* dataset which was captured with a Canon EOS 700D. The corresponding lenses have focal length 135 mm and 160 mm (in 35 mm equivalent) and approximate an orthographic camera. All results are computed on non-linear JPEG images. We intentionally did not remove gamma correction since dealing with non-linear intensities is one of the strengths of our technique.

5.2. Evaluation

To create a textureless target object we spray painted a bottle and an example sphere with brown paint such that they have a BRDF with a broad highlight[1], see Figure 6a. The shape of the bottle is rather uniform and can be recovered quite well as shown in Figure 7. Even the fine grooves are visible in the normals and the triangulated depth map. Our algorithm is also able to cope with differences in BRDF between the target and the reference sphere to a certain degree. We captured an additional dataset that contains the brown bottle (*bottle2*) and a white perfectly Lambertian sphere. We manually adjusted the albedo in the ap-

[1] The dataset is available at www.gris.informatik.tu-darmstadt.de/projects/mvps_by_example.



(a) (b) (c) (d)

Figure 6. Datasets with varying reflectance. (a-d) Cropped input images for the *bottle*, *diffuse owl*, *shiny owl*, and *spheres* datasets corresponding to the depth and normal maps shown in Section 5. We use the textured patterns in each scene to estimate camera pose.

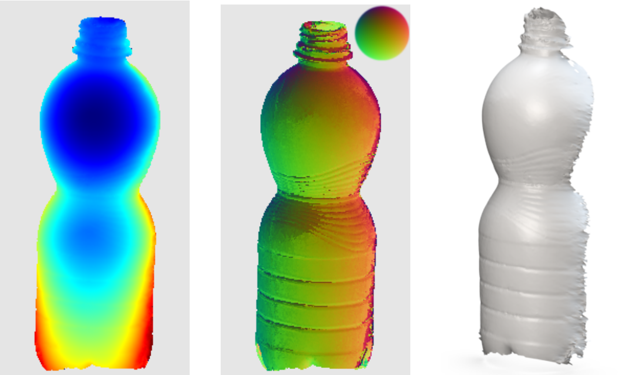


Figure 7. Results for the *bottle* dataset. Left to right: Colored depth map from blue (near) to red (far), the normal map, and a rendering of our triangulated geometry from a novel view.

pearance profiles of the bottle to approximate a white color. Note that this does not change the reflectance behavior and does in particular not change the (occurrence of) the specular highlight on the bottle. Figure 8 shows results that are only slightly degraded compared to the *bottle* dataset (see Figure 7) for which target and reference had the same reflectance. We also acquired a ground truth model for the *bottle* and *bottle2* datasets with a structured light scanner and registered it using an iterative closest point algorithm. Figure 9 shows two planes that cut through the ground truth and our depth maps. We observe that the deviations are less than 2.5 mm. This is at the scale of the alignment error, given that the camera was 2 m distant.

The *diffuse owl* is a 12 cm tall porcelain figurine which we spray painted with a diffuse green color to create a homogenous reflectance, see Figure 6b. The initialization in Figure 10 already provides good normals in many places, but our final result shows clear improvements especially at difficult regions such as the feet and around the eye. The rendering shows fine details and only some artifacts at depth discontinuities. After we captured the *diffuse owl* dataset, we applied a transparent varnish to the figurine which makes it appear glossy as shown in Figure 6c. This novel *shiny owl* dataset demonstrates our performance on non-diffuse surfaces. Even small details such as the feathers are clearly recognizable in Figure 11.

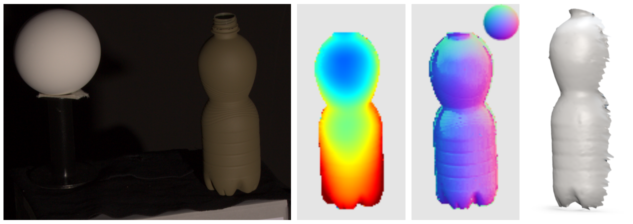


Figure 8. Matching different BRDFs. *Left to right*: An input image showing the diffuse white sphere next to the slightly shiny bottle, the recovered depth map (blue: near, red: far), the normal map, and a rendering from a novel view point.

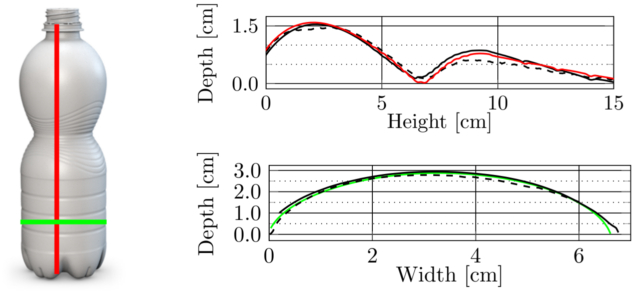
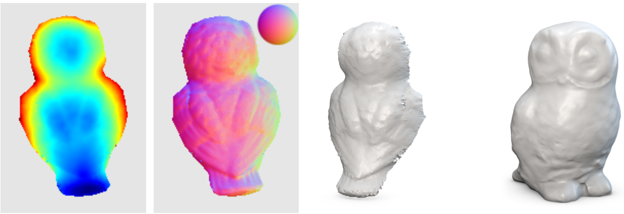


Figure 9. *Left*: Ground truth acquired from structured light scanning with horizontal (green) and vertical (red) profile lines. *Right*: The vertical (top) and horizontal (bottom) cuts through the ground truth (colored) and our depth map (black) show a deviation of less than 2.5 mm for the *bottle* (solid) and *bottle2* (dashed) datasets.



(a) (b) (c) (d)

Figure 11. (a-c) Results for the *shiny owl* dataset. Even for shiny surfaces, fine details can be recovered. (d) Novel view of a globally consistent model obtained by merging 17 depth maps of the *diffuse owl* dataset.

Integrating normal maps may result in globally deformed surfaces if it is not sufficiently constrained by depth information [16]. This can lead to problems if several views' ge-

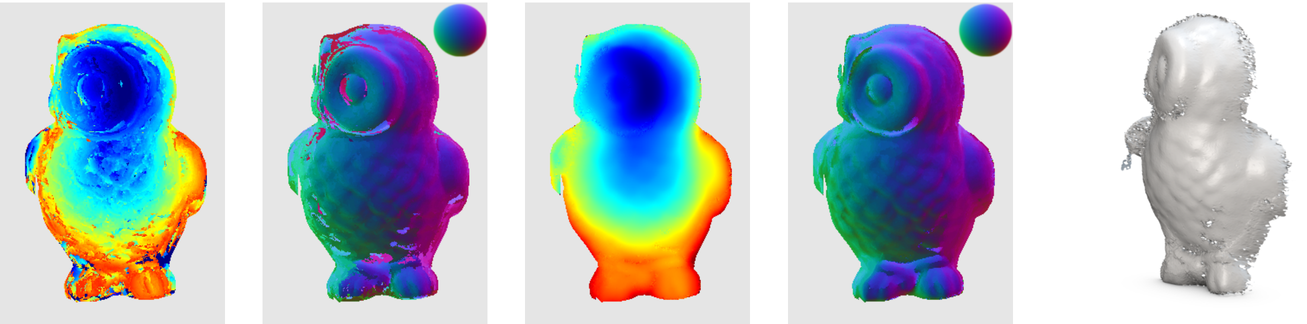


Figure 10. Improvement through optimization. *From left to right*: The initial depth and normal map for the *diffuse owl* dataset; our final depth and normal map after 50 iterations; the triangulated depth map rendered from a novel view.

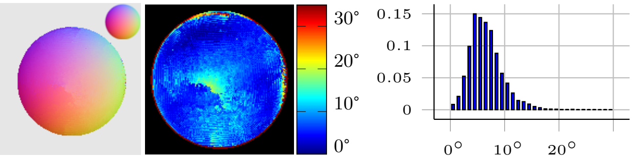
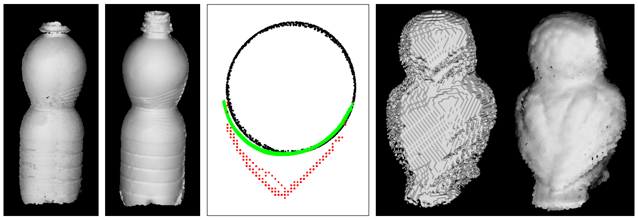


Figure 12. *Left*: Resulting normal map for the *spheres* dataset. *Middle*: The angular error compared to an ideal sphere. *Right*: Histogram over all angular errors below 30° for the *sphere*.

ometry is merged into a global model. Our integrated depth maps, however, are very consistent. Figure 11d shows a global mesh fused from 17 views. All depth and normal maps were projected to oriented 3D points and then processed using Poisson Surface Reconstruction [15].

To assess the maximal quality we can expect in practice, we use two transparent Christmas balls lacquered from the inside with acrylic paint, see Figure 6d. We use the left one as reference object and reconstruct the one on the right. This way we can quantitatively compare the reconstructed normals in Figure 12 against those of an ideal sphere whose position we obtain as described for the reference sphere. Small errors in that estimated position lead to a peak at 5° for the histogram of angular deviations in Figure 12. Although the target is not perfectly round and its reflectance does not completely match the reference due to varying thickness of the dye coating, the overall deviation is low. Most of the larger errors—besides at the boundaries—occur at the sphere center where the over-exposed highlight was observed most often.

Matching appearance profiles in a multi-view setting has also been studied by Treuille *et al*. [29]. Unfortunately, that work does not contain a quantitative evaluation that we could compare against. We reimplemented their technique and show the results in Figure 13. The *diffuse owl* dataset contains views from all directions. Voxel coloring produces a reasonable but discretized reconstruction. Detail information encoded in the normals is only accessible for rendering. In contrast, our energy formulation is continuous in depth and thus leads to a fundamentally different optimization problem. We provide a quantitative comparison with our reconstruction for the *bottle* where ground truth is available. This dataset contains only 14 cameras that observe the object mostly from the front. It demonstrates that our approach copes well with a restricted set of camera positions. The voxel reconstruction is not able to recover the true shape because the matching error is not very discriminative. In contrast, our approach enforces consistency of reconstructed normals and depth which provides a clear advantage.



(a) (b) (c) (d)

Figure 13. Comparison to Treuille *et al*. [29]. (a) The voxel-based reconstruction of the *bottle* rendered using point splatting. (b) Our reconstruction shown from the same view. (c) Geometry comparison: several horizontal slices through the *bottle* reconstructed with our approach (green), Treuille *et al*. [29] (red), and structured light (black) are plotted on top of each other. (d) The marching cubes reconstruction of the volume by Treuille *et al*. is blocky as shown for the *diffuse owl* dataset (left). The attached normals do not contribute to the geometry and can be only be used for shading (right).

6. Conclusion

In this paper we have shown that it is possible to reconstruct detailed geometry of objects observed from multiple views with challenging, unknown reflectance properties and lighting by matching with an example object. Our formulation is continuous in depth and operates directly on image intensities. In contrast to other methods, the final surface can therefore be optimized without referring to proxy geometry obtained from non-photometric techniques based on texture information or silhouettes. Representing the surface as depth maps instead of as a global model allows the use of well-understood image-based smoothness constraints and is easy to integrate with existing stereo approaches. Although we need a reference object with similar reflectance (the "example"), we believe that the generality that such an object offers in terms of unknown light setup and camera response are well worth the effort. Our results also show that the requirement of similar reflectance can be relaxed without sacrificing too much quality.

The computation times for a single view are quite high because we exhaustively match the per-pixel profiles against all reference profiles. In the future, we would like to speed up our prototypical implementation with GPU parallelization. The current formulation allows depth discontinuities but assigns them a large error. Thus, at boundaries and steep edges sometimes artifacts can occur. We would like to experiment with robust loss functions to address this in the future. Finally, it would be interesting to extend this technique to objects with mixed materials, *e.g.*, by introducing a second reference object with a different BRDF.

Acknowledgments: This work was supported in part by the European Commission's Seventh Framework Programme under grant agreements no. ICT-323567 (HARVEST4D) and no. ICT-611089 (CR-PLAY), as well as the DFG Emmy Noether fellowship GO 1752/3-1.

References

[1] S. Agarwal, K. Mierle, and Others. Ceres solver. code.google.com/p/ceres-solver. 4

[2] N. Alldrin, T. Zickler, and D. Kriegman. Photometric stereo with non-parametric and spatially-varying reflectance. In *CVPR*, 2008. 2

[3] J. T. Barron and J. Malik. Color constancy, intrinsic images, and shape estimation. In *ECCV*, 2012. 2

[4] H. Du, D. B. Goldman, and S. Seitz. Binocular photometric stereo. In *BMVC*, 2011. 2

[5] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. Seitz. Multi-view stereo for community photo collections. In *ICCV*, 2007. 2

[6] D. Goldman, B. Curless, A. Hertzmann, and S. Seitz. Shape and spatially-varying BRDFs from photometric stereo. In *ICCV*, 2005. 2

[7] H. Hayakawa. Photometric stereo under a light source with arbitrary motion. *JOSA A*, 11(11), 1994. 2

[8] C. Hernandez, G. Vogiatzis, and R. Cipolla. Multiview photometric stereo. *PAMI*, 2008. 1, 2

[9] A. Hertzmann and S. Seitz. Shape and materials by example: a photometric stereo approach. In *CVPR*, 2003. 2

[10] A. Hertzmann and S. Seitz. Example-based photometric stereo: shape reconstruction with general, varying BRDFs. *PAMI*, 27(8):1254–1264, 2005. 2

[11] T. Higo, Y. Matsushita, and K. Ikeuchi. Consensus photometric stereo. In *CVPR*, 2010. 2

[12] H. Jin, S. Soatto, and A. J. Yezzi. Multi-view stereo reconstruction of dense shape and complex appearance. *International Journal of Computer Vision*, 2005. 2

[13] M. K. Johnson and E. H. Adelson. Shape estimation in natural illumination. In *CVPR*, 2011. 2

[14] N. Joshi and D. Kriegman. Shape from varying illumination and viewpoint. In *ICCV*, 2007. 1, 2, 4

[15] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *SGP*, 2006. 7

[16] R. Klette and K. Schluens. Height data from gradient fields. In *SPIE Proc. Machine Vision Applications, Architectures, and Systems Integration V*, 1996. 6

[17] P.-Y. Laffont, A. Bousseau, S. Paris, F. Durand, and G. Drettakis. Coherent intrinsic images from photo collections. *ACM Transactions on Graphics*, 31, 2012. 2

[18] J. Lim, J. Ho, M.-H. Yang, and D. Kriegman. Passive photometric stereo from motion. In *ICCV*, 2005. 1, 2

[19] F. Lu, Y. Matsushita, I. Sato, T. Okabe, and Y. Sato. Uncalibrated photometric stereo for unknown isotropic reflectances. In *CVPR*, 2013. 2

[20] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3d geometry. In *ACM SIGGRAPH*, 2005. 2

[21] G. Oxholm and K. Nishino. Shape and reflectance from natural illumination. In *ECCV*, 2012. 2

[22] G. Oxholm and K. Nishino. Multiview shape and reflectance from natural illumination. In *CVPR*, 2014. 2

[23] T. Papadhimitri and P. Favaro. A new perspective on uncalibrated photometric stereo. In *CVPR*, 2013. 2

[24] J. Park, S. N. Sinha, Y. Matsushita, Y.-W. Tai, and I. S. Kweon. Multiview photometric stereo using planar mesh parameterization. In *ICCV*, 2013. 1, 2, 4

[25] I. Sato, T. Okabe, Q. Yu, and Y. Sato. Shape reconstruction based on similarity in radiance changes under varying illumination. In *ICCV*, 2007. 2

[26] B. Shi, Y. Matsushita, Y. Wei, C. Xu, and P. Tan. Self-calibrating photometric stereo. In *CVPR*, 2010. 2

[27] B. Shi, P. Tan, Y. Matsushita, and K. Ikeuchi. Elevation angle from reflectance monotonicity: Photometric stereo for general isotropic reflectances. In *ECCV*, 2012. 2

[28] W. M. Silver. Determining shape and reflectance using multiple images. Master's thesis, MIT, 1980. 2

[29] A. Treuille, A. Hertzmann, and S. Seitz. Example-based stereo with general BRDFs. In *ECCV*, 2004. 1, 2, 3, 7

[30] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1), 1980. 2

[31] C. Wu. Towards linear-time incremental structure from motion. In *3DV*, 2013. 2

[32] Y. Yoshiyasu and N. Yamazaki. Topology-adaptive multiview photometric stereo. In *CVPR*, 2011. 1, 2

[33] Q. Zhang, M. Ye, R. Yang, Y. Matsushita, B. Wilburn, and H. Yu. Edge-preserving photometric stereo via depth fusion. In *CVPR*, 2012. 1, 2

[34] Z. Zhou, Z. Wu, and P. Tan. Multi-view photometric stereo with spatially varying isotropic materials. In *CVPR*, 2013. 2