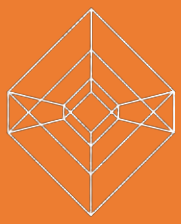


HARVEST4D

HARVESTING DYNAMIC 3D WORLDS FROM COMMODITY SENSOR CLOUDS

Publications for Task 6.4

Deliverable 6.41

Date:	21.6.2016
Grant Agreement number:	EU 323567
Project acronym:	HARVEST4D
Project title:	Harvesting Dynamic 3D Worlds from Commodity Sensor Clouds

Document Information

Deliverable number	D6.41
Deliverable name	Publications for Task 6.4
Version	1.0
Date	2016-06-21
WP Number	6
Lead Beneficiary	CNR
Nature	R
Dissemination level	PU
Status	Final
Author(s)	CNR

Revision History

Rev.	Date	Author	Org.	Description
0.1	21/6/16	G.Palma, P.Cignoni	CNR	First Draft
0.2				

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY

1.1 INTRODUCTION

This deliverable describes the publications that resulted from Task 6.4, and how they fit into the work plan of the project.

The objective of Task 6.4 is to detect actual changes in the reflectance of the surface of scenes acquired multiple times. This task is very challenging because we need solutions that deal not only with geometric changes of the objects but also with variations in the lighting conditions. The final purpose is to design new techniques able to reconstruct the evolution in time of the surface reflection properties of the sampled surfaces with respect to different acquisitions. This would allow, for example, monitoring visual conditions of Cultural Heritage artefacts that can be affected by several degradation process, like rusting, weathering or damages.

There are two publications that are mainly attributable to Task 6.4, and these can be found in the appendix of this deliverable.

1.2 PUBLICATIONS

The following publications can be found in the appendix:

- Gianpaolo Palma, Francesco Banterle and Paolo Cignoni.
  Temporal Appearance Change Detection using Multi-View Image Acquisition.
  Technical report ISTI CNR. June 2016.

- Jens Ackermann, Fabian Langguth, Simon Fuhrmann, Arjan Kuijper and Micheal Goesele.
  Multi-View Photometric Stereo by Example.
  In: Proceedings of the International Conference on 3D Vision, 2014.

2 DESCRIPTION OF PUBLICATIONS

2.1 OVERVIEW

The objective of this task is to develop new algorithms to detect appearance changes of a surface. The final goal is separating the changed parts of the reflectance in order to facilitate applications like the temporal monitoring of an artefact or estimating a more complete appearance model. In this context, the following approaches propose different solutions working on similar input data: multi-view photo datasets. Specifically, the work [Palma et al. 2016] allows detecting appearance changes over time using an explicit model of surface reflectance, the Surface Light Field. Contrary, the work [Ackermann et al. 2014] allows for change detection with respect to illumination conditions. This is possible by computing a globally consistent model of geometry and material properties for an object to be reconstructed. The input is composed by a set of images with an additional reference object acquired with varying light and camera positions.

2.2 TEMPORAL APPEARANCE CHANGE DETECTION USING MULTI-VIEW IMAGE ACQUISITION

This paper presents a novel solution for the automatic detection of temporal appearance changes on a surface using the comparison of an explicit reflectance model. Starting from two sets of multi-view photos acquired at different times, the algorithm computes the 3D model by multi-view 3D reconstruction and the per-vertex Surface Light Field (SLF) for each time independently. A Surface Light Field approximates the reflectance function of the object and allows estimating the appearance from different viewing directions with respect to the lighting environment that was in effect during the acquisition of the input photos. The SLF is estimated with a new algorithm that extends the method presented in Task 7.2 by a more robust separation between diffuse color and residual reflectance effects. Then, the SLFs from different time instants are compared with a weighted approach taking into account small lighting variations and small misalignments in the color-to-geometry projection. In particular, we compute two different change fields: the change field of the diffuse color and the change field of the residual reflectance effects. On a dataset with synthetic changes (Figure 1), our results show several interesting features that can give the user cues about the areas affected by changes. These results can serve as a good starting point for further research to improve the accuracy of the detection, especially by removing false-positive case.

Figure 1. Comparison of the change fields with the segmentation on the input image of the areas affected by the changes. (Left) Input image. The changed regions are highlighted in blue. (Center) Change field of the diffuse color mapped on a color ramp from blue to red. (Right) Change field of the residual SLF effects mapped on a color ramp from blue to red.

2.3 MULTI-VIEW PHOTOMETRIC STEREO BY EXAMPLE

The approach of the publication *Multi-View Photometric Stereo by Example* [Ackermann et al. 2014] facilitates reconstruction of an object of interest by means of an additional example object. The method produces a globally consistent model of the surface geometry enriched with additional material properties in form of a Bidirectional Reflectance Distribution Function (BRDF).

As the name implies, the approach requires multiple views of the scene as input, where both, the object of interest and the reference object, must be visible in each of the images (Figure 2). The method exploits the similarity between target and reference object by developing a consistency measure between both based on orientation consistency. This allows a reconstruction with as little additional information as possible. The method especially supports the challenging case of reconstruction without knowing the object BRDF or scene illumination. Furthermore, the capture conditions can be uncontrolled apart from the reference object since the method supports changing camera and light positions between images. Because normals can be recovered more reliably than depth, the object surfaces are represented by means of both a depth and a normal map. These two maps are jointly optimized which allows formulating constraints on the depth which consider surface orientation. Using both map types leads to increased robustness and accuracy.

Figure 2. Cropped example input image of the *shiny owl* multi-view dataset showing the object of interest on the left and the used reference object on the right side.

Beneficially, the approach works on uncontrolled image intensities only and does not require radiometric camera calibration. Moreover, neither a visual hull nor stereo reconstructions for bootstrapping are necessary. The results show that the method works on textureless real world objects. It is even possible to create globally consistent models in presence of challenging specular reflectance (Figure 3).

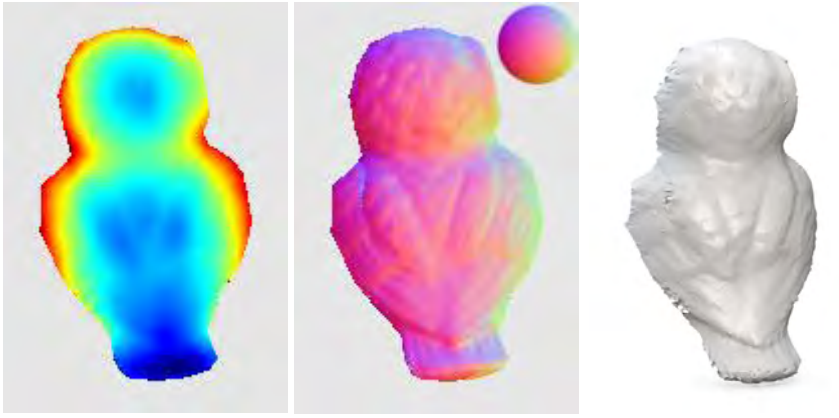


Figure 3. Results for the *shiny owl* dataset. Left to right: Colored depth map from blue (near) to red (far), the normal map and a rendering of a novel view via triangulated geometry.

With regard to scene changes, the method not only supports changing acquisition conditions, such as varying camera and lighting positions, but also allows the reconstruction of a globally consistent model comprised of estimated geometry and material properties in form of a BRDF. This complex surface representation facilitates calculation of changes in illumination by determination of illumination conditions for each individual image.

3 REFERENCES

- Gianpaolo Palma, Francesco Banterle and Paolo Cignoni.
  Temporal Appearance Change Detection using Multi-view Image Acquisition.
  Technical report ISTI CNR. June 2016.
- Jens Ackermann, Fabian Langguth, Simon Fuhrmann, Arjan Kuijper and Micheal Goesele.
  Multi-View Photometric Stereo by Example.
  In: Proceedings of the International Conference on 3D Vision, 2014.

4 APPENDIX

The following pages contain all the publications that are directly associated with this deliverable. Other publications referenced in this deliverable can be found in the public Harvest4D webpage (for already published papers), or in the restricted section of the webpage (for papers under submission, conditionally accepted papers, etc.).

Temporal Appearance Change Detection using Multi-View Image Acquisition

Gianpaolo Palma[1], Francesco Banterle[1], Paolo Cignoni[1]

[1]Visual Computing Lab - ISTI - CNR, Pisa, Italy

Abstract

Appearance change detection is a very important task for applications monitoring the degradation process of a surface. This is especially true in Cultural Heritage (CH), where the main goal is to control the preservation condition of an artifact. We propose an automatic solution based on the estimation of an explicit parametric reflectance model that can help the user in the detection of the regions that are affected by appearance changes. The idea is to acquire multi-view photo datasets at different times and to compute the 3D model and the Surface Light Field (SLF) of the object for each acquisition. Then, we compare the SLF in the time using a weighting scheme, which takes account of small lighting variations and small misalignments. The obtained results give several cues on the changed areas. In addition, we believe that these can be used as good starting point for further investigations.

1. Introduction

The acquisition and estimation of the surface appearance of a real object is a fundamental step towards the realistic rendering of its 3D model. This topic was extensively studies with the proposal of several robust solutions that are different for the type of acquisition setup (a totally controlled setup, like a dark room [LKG*03], or a general and uncontrolled lighting environment [PCDS12]) and for the approximation of the appearance that we want to reproduce (Spatially Varying Bidirectional Reflectance Distribution Function [LKG*03] [PCDS12], Surface Light Field [PDCS13] or simple apparent color [CCCS08]). Given these solutions, it is now possible to think about new algorithms to detect the appearance changes over time of an object in order to understand the type of changes and where they have happened. A potential application of these algorithms is the monitoring of the surface condition and evolution of a Cultural Heritage artifact that can be altered by several degradations process such as opacification (lose of shininess), rusting, weathering, lose of small pieces of painting, dust accumulation, human damages, etc.

In this paper, we present a novel solution to detect automatically the area of the objects that suffer of some type of appearance change using dataset of multi-view images acquired in different times and with similar lighting condition. The method uses the same images to compute both a 3D model of the object and an approximated appearance model as the Surface Light Field for each time step independently. A Surface Light Field is an approximation of the reflectance function of the object that allows the rendering of the appearance from different view directions taking fixed the lighting environment. This is the same environment used during the acquisition of the input photos. Subsequently, the computed Surface Light Fields are compared taking into account small lighting changes and misalignments. The main contributions of our method are:

- a procedure to estimate the Surface Light Field that extends the solution proposed in [PDCS13] with a more robust separation between the diffuse color and the other residual reflectance effects;
- a comparison procedure of two Surface Light Fields of the same object acquired in different times that takes account of small changes due to different lighting conditions between the two captures and to small misalignment in the projection of the color data over the mesh from the input photos and between the 3D models.

2. Related Work

The general problem of the change detection has been intensively studied in the Computer Vision community with the goal of detecting regions of change in images of the same scene taken at different times. Early methods are based on simple per-pixel intensity differences, choosing empirical segmentation thresholds [Ros02] [RI03]. There are several methods that are closely related to simple differencing, as the change vector analysis [BP02], often used for multispectral images, or the image rationing that uses the ratio, instead of the difference, between the pixel intensities [SIN89]. Other approaches are based on a statistical modeling of the problem, like the test of the null hypothesis that a pixel is a change or not assuming a Gaussian random variable with zero mean [AK95] or the use of Probabilistic Mixture Models [BFY00]. More sophisticated algorithms exploit the relationships between close pixels both in space and time, and they fit the intensity values of each local block to a polynomial function of the pixel coordinates [HNR84] or use an autoregressive process over the time [Cli03]. Other techniques are based on a shading model to produce illumination-invariant algorithm [LL02]. Typically, the output of these change detection algorithms is a mask, where decisions are made independently for each pixel. This mask is noisy with isolated change pixels, holes in the middle of connected change components, and jagged boundaries. Since changes in real image sequences often arise from the appearance or motion of solid objects with continuous and differentiable boundaries, most change detection algorithms try to conform the change mask using either standard binary image processing operations [Str00] or concepts from Markov Random Fields (MRFs) [KV02].

Most change detection approaches start by estimating the geometric mapping (the geometry of the scene and the camera parameters) among images due to the viewpoint change. A common assumption in remote sensing and CCTV scenarios is to assume static either purely rotating cameras or planar scene, while in the other case more general multi-view stereo methods are employed [TBP11]. Recently, Sakurata et al. [SOD13] proposed a change detection approach that avoids the explicit determination of geometry of the scene by integrating for change over depth uncertainties.

Alternative approaches show the effectiveness of using deep learned features for this change detection task. Sakurata et al. [SO15] proposed a solution for the detection in a pair of vehicular and omnidirectional images. This uses a convolutional neural network to compute a rough segmentation of the changed regions in combination with a superpixel segmentation to refine the boundaries of the changed regions. Stent et al. [SGSC15] described a system for the detection of changes in multiple views images of a tunnel surface. They proposed to use a two-channel convolutional neural network for detecting changes as hairline cracks, water ingress, and other surface damages. They trained the network on synthetically generated examples.

Feng et al. [FTZ*15] proposed a solutions based on the acquisition of multiple images with multiple illumination. They formulated a fine-grained change detection as a joint optimization problem of three factors: normal-aware lighting difference; camera geometry correction flow; real scene change mask. They proposed to solve the three factors in a coarse-to-fine manner and achieve reliable change decision by rank minimization. Similarly, Stent el al. [SGSC16] introduced a precise deterministic approach for pixel-wise change detection in pair of images of a scene of interest taken over time with similar illumination. The approach compensates for the three most common sources of variation: viewpoint variation due to camera motion between images, photometric variation due to lighting differences, and changes in image resolution/focal settings.

Our method is the first algorithm that tries to detect the appearance changes by comparison over the time of an explicit parametric reflectance model.

3. Algorithm Overview

The main goal of the proposed method is to detect the regions of the object affected by appearance changes in the time. Starting from two multi-view image datasets of the object acquired at different times $A = \{I_i \dots I_n\}$ and $B = \{I_i \dots I_m\}$, the main steps of the algorithm are the reconstruction of the 3D models of the object of interest using the input images with a multi-view dense reconstruction algorithm, the estimation for each time step of the Surface Light Field (SLF), and finally the comparison of the estimated SLF to compute a change field over the surface. We assume that each image dataset was acquired in fixed lighting condition (the lighting condition does not change during acquisition), and that the lighting environment remained similar between the two capture sessions with some small differences.

The SLF is estimated for each vertex of the 3D model using an extension of the method proposed in [PDCS13] based on a more robust separation of the diffuse color from the other reflectance effects. The main idea in [PDCS13] is to separate the estimation of the diffuse component of the surface appearance from the other view dependent lighting effects. The first one is modeled as a simple RGB color while the residual effect as a linear combination of Hemispherical Harmonics [GKPB04]. This separation avoids rendering artifacts due to the fitting and interpolation process of the hemispherical functions. The final color of a point p is given as:

$$SLF(p,s,t) = D(p) + \sum_{i=0}^{n} x_i(p) h_i(s,t), \qquad (1)$$

where (s,t) are the spherical coordinate of the view vector $\vec{v}$ in the local tangent space of the point p, $D(p)$ is the diffuse

color, and $x_i(p)$ is a coefficient associated to the used basis hemispherical functions $h_i(s,t)$.

The final comparison is done independently for each component of the SLF using a weighting scheme that takes account of small lighting variations during the two acquisitions and of small misalignments.

4. Image Preprocessing and 3D reconstruction

The first step of the method is the linearization of the two image datasets using an estimation of the Camera Response Function (CRF). We also want to reduce all the image to the same time exposure and camera aperture in order to have comparable colors among the different views. For this task we acquire a series of images of a Macbeth chart at different exposure times to estimate the CRF using the Mitsunaga-Nayar method [MN99].

The linearization and normalization of the i-th image, I_i, is defined as

$$I_i' = \text{ApplyCRF}\left(\text{RemoveCRF}\left(I_i\right)\frac{F_i^2}{t_i}\,\frac{t_{ref}}{F_{ref}^2}\right), \qquad (2)$$

where I_i' is the processed image, t_i and t_{ref} are, respectively, the exposure time of I_i and of the reference, and F_i and F_{ref} are, respectively, the camera aperture of I_i and of the reference.

Note that, we assume that the same camera was used to acquire both datasets. In the case this assumption is not true, we need an extra photometric calibration in order to put all the images in a common color space and gamut by estimation of a color transformation matrix for each different camera.

The next step is the computation of a triangular mesh for each time step independently using the input multi-view images (see Figure 1). For each time step, we export also the camera parameters (intrinsic and extrinsic parameters) of the input photos. These are needed in the following step for projecting the color info over the mesh and for estimating the SLF.

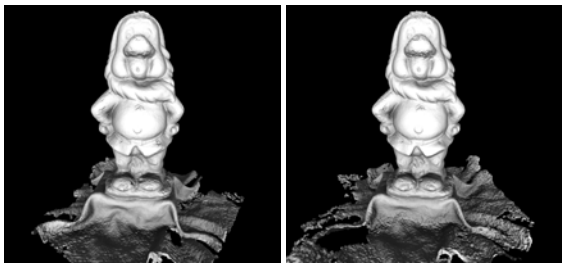


Figure 1: *3D models obtained by the input images of the two datasets.*

5. SLF Estimation

We estimate the SLF for each vertex of a mesh. The values inside the triangles are obtained using the barycentric interpolation. Anyway, the method can effortlessly extended for the case of a mesh with a texture parameterization.

First, we collect the color samples projected by each photo on the vertices of the mesh as

$$C_p = \{I_i'(x,y)|(x,y) = M_i p\}, \qquad (3)$$

where M_i is the model-view-projection matrix computed using the camera parameters of the photo I_i returned by the 3D reconstruction. For this task, we use a straightforward projection on the GPU. At this step, for each color sample, we compute the distance in pixels from the nearest depth discontinues $b_i(p)$ to penalize wrong color samples due to small misalignments. For computing this weight, $b_j(p)$, we render the model using the estimated camera matrix. Then, we extract the edges from the depth map using the image Laplacian operator, and we detect the most valuable borders using the 0.95 percentile of the histogram of the edge map. Finally, we compute the distance field from these borders using a GPU jump flooding algorithm [RT06].

The main contribution in the estimation of the SLF is a new approach for the robust separation between the diffuse color and the other residual reflectance effects. The goal is to obtain a diffuse color that is free from residual specularity defects that other color blending solutions can create; see differences from [CCCS08] and [PCDS12] in Figure 2.

Figure 2: *Comparison of the diffuse color estimated by our method (Left) with two state of the art solutions: (Right) [CCCS08]; (Center) [PCDS12].*

The algorithm starts by computing the diffuse component of the SLF using the solution proposed in [PCDS12]. In more details, we estimate a rough estimation of the lighting environment by computing a threshold for each vertex equals to the sum of mean and absolute deviation of the luminance of the samples projected on the vertex. Then, we project the samples with luminance above this threshold on a environment map using the specular mirror direction of the

view vector, accumulating in the environment map the differences from the threshold. This environment map, normalized in the range $[0,1]$, is used to computed another weight for each sample, the specular weight $s_i(p)$. This gives a probability of the sample to show a specular behavior. We compute the specular by sampling the rough computed environment map in a cone of direction along the specular mirror direction (see [PCDS12] for more details). Finally, we blend the color samples projected on the vertex using a simple weighted mean:

$$D(p) = \frac{\sum_{c_i \in C_p} c_i(p)\, w_i(p)}{\sum_{c_i \in C_p} w_i(p)} \tag{4}$$

where the weight $w_i(p) = \max(b_j(p)/64, 1.0)(1.0 - s_j(p))(1.0 - \mathrm{lum}(c_i)(p))$ is defined as product of three measures: the border weight b_j normalized in the range $[0,1]$ using a normalized threshold of 64 pixels that penalizes wrong colors due to small misalignments between the photo and the geometry; one minus the specular weight s_j to give more weight to the samples that have a lower probability to exhibit a specular reflectance behavior; one minus the luminance of the color sample to give more weight to the samples with a lower luminance. We apply this procedure to compute two different versions of the diffuse color: the first one, $D_{blend}(p)$, uses all the color samples projected on p (first column in Figure 4); the second one $D_{min}(p)$ uses only the five color samples with the lowest luminance value (second column in Figure 4). There are some important differences between the two versions. D_{blend} shows a smooth color variation over the surface with a more uniform color but it presents also very bright areas due to an higher persistent of a specularity in these regions in the input images (second row in Figure 4). D_{min} shows a color that is nearer to the real diffuse color without residual specularity but with some abrupt color differences over the surface (third row in Figure 4).

To estimate a more consistent diffuse color, we try to transfer the smoothness of the color variation from D_{blend} to D_{min} adapting two image processing algorithms to a 3D mesh: the color histogram matching and the Poisson image editing [PGB03]. As first step, we compute the histogram matching between D_{blend} and D_{min} to transfer the luminance distribution from D_{min} and D_{blend}. We simply adapt the algorithm by computing the Cumulative Distribution Function (CDF) of the luminance of the per-vertex color of the two versions. Then, we compute the transformation for matching the CDF of D_{blend} to the CDF of D_{min}. Finally, we apply this transformation to D_{blend} to obtain a new version of the diffuse color D_{histo} (third column in Figure 4). This new color shows better luminance level than D_{blend} but with the same residual specularity. The next step is to transfer the local gradient from D_{histo} on D_{min} using a modified Poisson Image Editing approach [PGB03]. The final goal is to correct the abrupt color changes in D_{min}. The general idea is to select as target regions the vertices that have an high local color gradient in D_{min} and a very low color gradient in D_{histo}. For the computation of the per-vertex color gradient $\nabla D_{histo}(p)$ and $\nabla D_{min}(p)$, respectively for D_{histo} and D_{min}, we use the average of the color differences from the 1-ring neighbor vertices N_p defined as

$$\nabla D_{histo}(p) = \frac{\sum_{q \in N_p} \|D_{histo}(p) - D_{histo}(q)\|_2}{|N_p|} \qquad \nabla D_{min}(p) = \frac{\sum_{q \in N_p} \|D_{min}(p) - D_{min}(q)\|_2}{|N_p|}. \tag{5}$$

The target region is defined as the set of vertices $\Omega = \{p \mid \nabla D_{histo}(p) < \beta \wedge \nabla D_{min}(p) > D_{histo}(p)\}$ and the boundary set is defined as $\partial\Omega = \{p \notin \Omega \mid \exists q \in S : p \in N_q\}$. The final solution for the vertices in Ω is computed by solving the following system of linear equation for each color channel independently to update the color in D_{min}:

$$\forall p \in \Omega \qquad |N_p| D_{min}(p) - \sum_{q \in N_p \cap \Omega} D_{min}(q) = \sum_{q \in N_p \cap \partial\Omega} D_{min}(q) + \sum_{q \in N_p} v_{pq}, \tag{6}$$

where v_{pq} is defined as

$$v_{pq} = (D_{histo}(p) - D_{histo}(q)). \tag{7}$$

We use an iterative approach. At each iteration, we compute $\nabla D_{min}(p)$, the set Ω and $\partial\Omega$ and we solve the system in Equation 7 to update D_{min}. We stop this process when the set Ω has less than 20 elements or when we reach 20 iterations. In order to avoid the creation of the defects in Figure 3, which are usual near to regions with high gradient in D_{histo}, we force the gradient differences v_{pq} to zero when $\nabla D_{histo} \geq \beta$. Finally, we need to transfer the gradient also between the regions with high color gradient in D_{histo}, defined by the vertices in $\Omega' = \{p \mid \nabla D_{histo}(p) \geq \beta\}$, using the same formulation in Equation 7. This last iteration allows us to create a more smooth color variation in the neighborhood of these areas. The final obtained color $D_{poisson}$ merges the good features of D_{histo} and D_{min} (fourth column in Figure 4): a smooth color over the surface without abrupt wrong local variation that reduces as soon as possible the residual specularity. In all our experiments we use the threshold $\beta = 15$ with RGB color defined in the range $[0,255]$.

Starting from the diffuse color $D_{poisson}$, we can model the view-dependent residual reflectance effects as combination of Hemipherical Harmonics using the same procedure describe in [PDCS13]. We retrieve the set of color samples $S(p) = \{c_i \in C(p) | \mathrm{lum}(c_i) > \mathrm{lum}(D_{poisson}(p))\}$ that have a positive luminance residual from the diffuse color and we solve a system of linear equations $Ax = b$. In this system, A is an $m \times n$ matrix that for each row, one for each sample in $S(p)$, contains the values of the Hemispherical Functions computed for the view direction of the sample, x is

Figure 3: *(Left) An artifact created by the poisson color correction; the black of the eye outline is propagated over the face. (Right) Improved version of the color imposing the gradient equals to zero on the boundary near to a big color variation. In this case, the color around the eye is sharper.*

the vector of the n coefficients to estimate and b is the vector with the luminance difference from the diffuse color. To solve the overdetermined system we use a Weighted Singular Value Decomposition (SVD) using a per-sample weight $w_i'(p) = \max(b_j(p)/64\ ,\ 1.0)s_j(p)\,\mathrm{lum}(p)$. In general the samples in S cover only a small part of the visible hemisphere. To avoid that the fitting procedure creates artifacts (banding and ringing effects) in the not sampled areas, we add some virtual samples, uniformly distributed in the uncovered regions, with a residual color equal to zero. The number of these samples depends on the maximum order of Hemispherical Harmonic functions used for the fittings (the higher is the order, the higher is the number of samples) and they are distributed with a Poisson-Disk pruning strategy with respect to the existing samples. The final models of the two time steps are:

$$SLF^A(p,s,t) = D^A_{poisson}(p) + \sum_{i=0}^{n} x_i^A(p) h_i(s,t)$$

$$SLF^B(p,s,t) = D^B_{poisson}(p) + \sum_{i=0}^{n} x_i^B(p) h_i(s,t)$$

6. SLF Comparison

Given SLF_A and SLF_B for the two time steps, the final task is to compare them. The idea is to compute two different change fields: the differences between the diffuse colors δD and the differences of the residual effects δHS. For this task, we need to geometrically align the two 3D meshes because the 3D reconstruction from images returns models in different reference systems and with different scales. We use a simple method based on the manual picking of some correspondences between the models to compute an initial rough transformation that is refined with the ICP algorithm [BM92]. An alternative is to use solutions based on the automatic detection of the correspondences between the models [MDS15].

The change fields are computed for each vertex of the models. In more details, for each vertex p of the acquisition A, the algorithm looks for the nearest point q in the model of acquisition B, it computes the SLF components for the point q using the barycentric interpolation of SLF of the vertices of the face that contains q and it computes the differences δD and δHS:

$$\delta D(p,q) = w_{diff}(p,q) \left\| D^A_{poisson}(p) - D^B_{poisson}(q) \right\|_2 \quad (8)$$

$$\delta HS(p,q) = w_{diff}(p,q) \sqrt{\sum_{i=0}^{n} (x_i^A(p) - x_i^B(q))^2}. \quad (9)$$

The same is done for the vertex of the model of the acquisition B. In this procedure, we discard all the pairs of point that are too distance, that is $\|p-q\|$ is above $1/1000$ of the bounding box of the mesh. The distance between the diffuse colors is computed in CIELAB color space assuming that the input color are in the color space sRGB and using the delta function defined in [SWD05]. Both the differences δD and δHS are weighted with a function $w_{diff}(p,q)$ that is the product of two Gaussians:

$$w_{diff}(p,q) = e^{-\frac{\delta s(p,q)^2}{0.05}} e^{-\frac{((\max(0.9\,,NCC(p,q))-0.9)/0.1)^2}{0.3}}, \quad (10)$$

where $\delta s(p,q)$ is a term that takes account of small lighting variations between the two captures, and $NCC(p,q)$ is a term that takes account of small misalignments in the projection of the color data over the 3D geometries.

The term $\delta s(p,q) = |s(p) - s(q)|$ is computed as difference between the per-vertex shading contribution of the lighting environment of each capture. The shading contribution is obtained as convolution between the visibility function of the vertex and an approximation of the lighting environment. The first task is to estimate the visibility function of each vertex $V(p) : \vec{\omega} \in \Omega \rightarrow 0,1$ to take account for effects of self-occlusion and self-shadowing. We precompute a spherical harmonics approximation with 36 coefficients using a simple ray casting of 256 rays per vertex:

$$\tilde{V}(\vec{\omega}, p) = \sum_{l=0}^{5} \sum_{m=-l}^{l} k^{(l)(m)}(p)\, Y^{(l)(m)}(\vec{\omega}). \quad (11)$$

Then, we estimate an approximation of the real lighting conditions by selection of the color samples to reproject and accumulate on a environment map. In particular, for all the color samples c_i with luminance above the per-vertex diffuse color, we reproject their difference from the diffuse color along the specular mirror direction $\vec{r}_i$ of the view vector $\vec{v}_i$, and we accumulate the value $x_i(p) = (\mathrm{lum}\, c_i(p) - \mathrm{lum}(D_{poisson}(p))\tilde{V}(\vec{r}_i, p)$ along this direction in the environment map. The obtained environment maps are normalized in the range $[0,1]$ and approximated with 36 coefficients of Spherical harmonics (Figure 5) such that:

$$\mathrm{EnvMap}(\vec{\omega}) = \sum_{l=0}^{5} \sum_{m=-l}^{l} g^{(l)(m)}\, Y^{(l)(m)}(\vec{\omega})). \quad (12)$$

The per-vertex shading contribution is computed as product

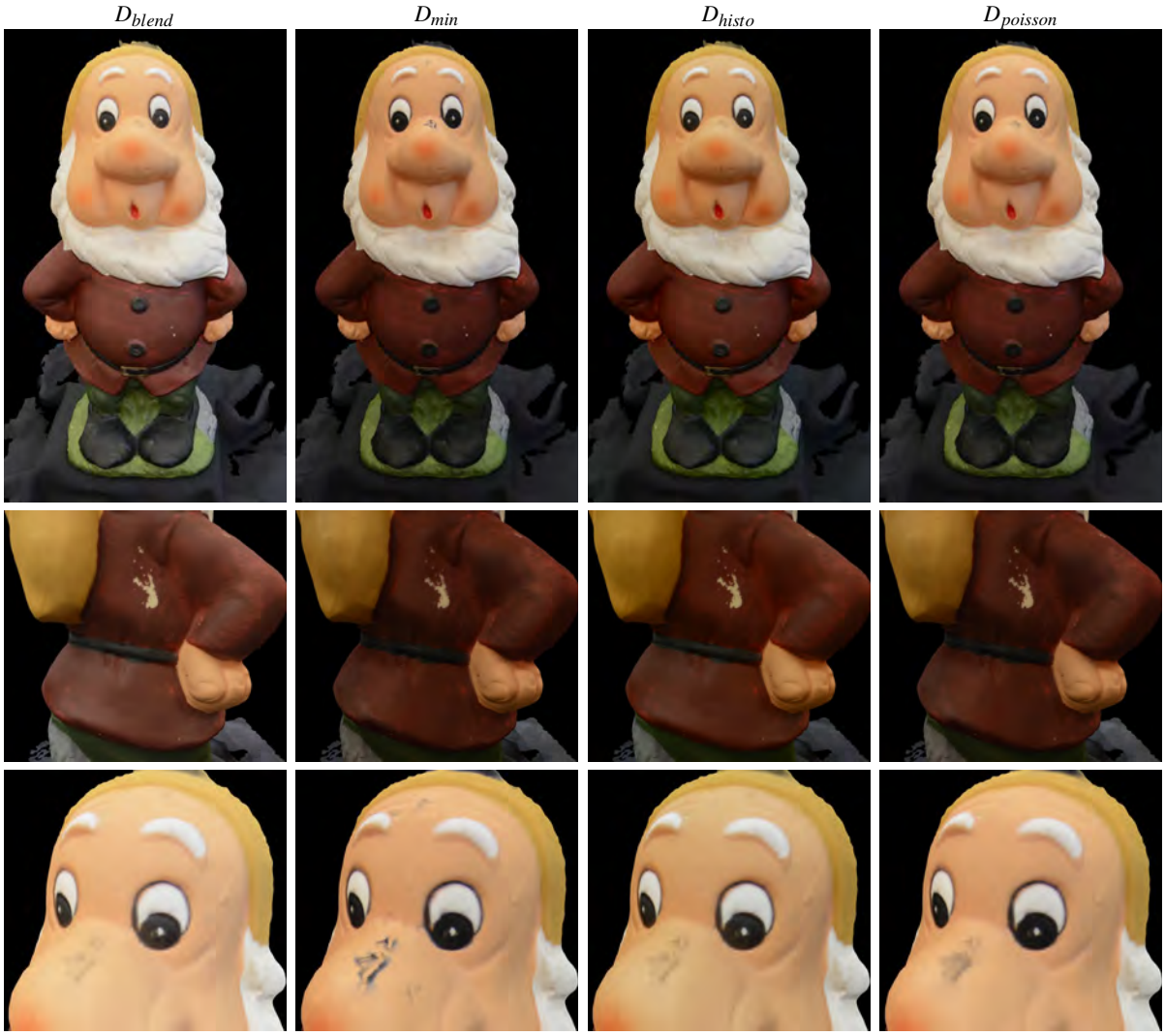


Figure 4: *A comparison of the different versions of diffuse color computed by our method. The second row shows critical cases for the color D_{blend}. The third row shows critical cases for the color D_{min}.*

of the Spherical harmonics coefficients of the visibility function and of the approximated environment map such that:

$$s(p) = \sum_{l=0}^{5} \sum_{m=-l}^{l} k^{(l)(m)}(p)\, g^{(l)(m)}. \quad (13)$$

Figure 6 shows the per-vertex shading contribution rendered in gray-scale colors for the two times.

The term $NCC(p,q)$ is the Normalized Cross Correlation between a patch around p and a patch around its close point in the other time q. In details, given a vertex p and its 1-ring neighbor vertices N_p defined by the local triangulation, we retrieve the closest points q and N_q in the other time, respectively for p and the points in N_p, we compute the means $\bar{p}$ $\bar{q}$ and the standard deviations σ_p σ_q of the sets $N_p \cup p$ and $N_q \cup q$ and we compute the NCC:

$$NCC(p,q) = \frac{1}{|N_p|\sigma_p\sigma_q} \sum_{p_i \in N_p, q_i \in N_q} ((p_i - \bar{p})(q_i - \bar{q})) \quad (14)$$

Figure 7 shows a color mapping of the NCC value in the range $[0.9, 1]$.

Figures 10 and 11 show a color mapping of the two change fields from different point of views. Figure 8 shows the differences in the computation of the change fields with and without the weighting function $w_{diff}(p,q)$.

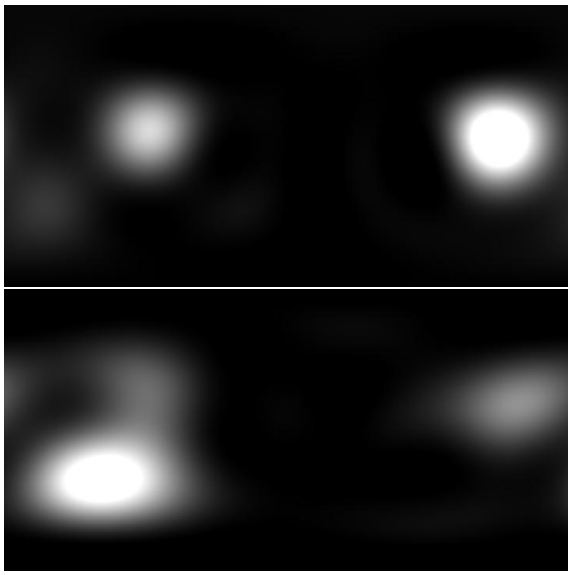


Figure 5: *Environment maps of the estimated lighting conditions used during the acquisition of the dataset A (Top) and the dataset B (Bottom).*

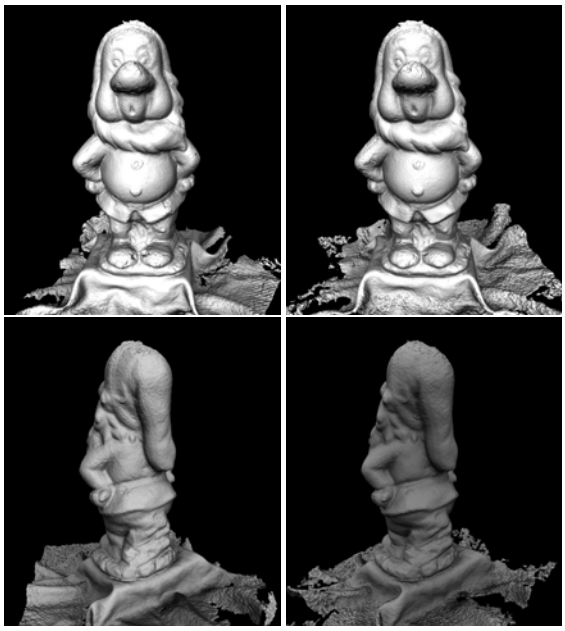


Figure 6: *Per-vertex shading contribution for dataset A (left) and dataset B (right).*

7. Results

To test the method, we created a time-varying reflectance dataset by introducing some synthetic changes on the surface of an object. We took a ceramic small statue of a dwarf, which is characterized by different types of specularity, and

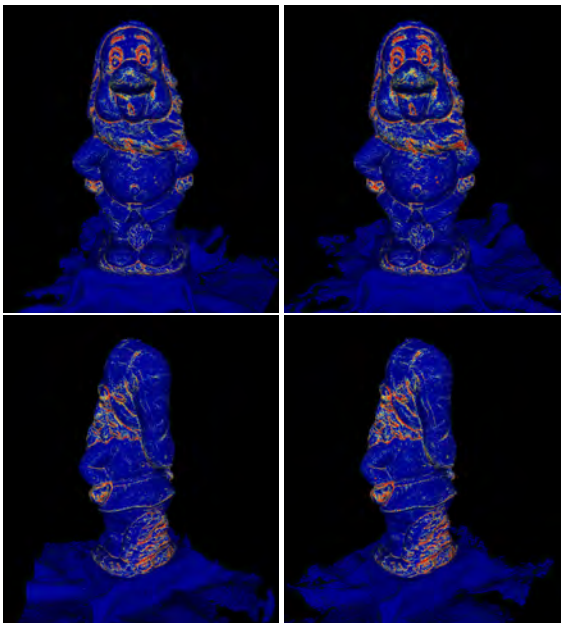


Figure 7: *Per-vertex NCC for dataset A (left) and dataset B (right) mapped on a color ramp from blue ($NCC = 0.9$) to red ($NCC = 1.0$).*

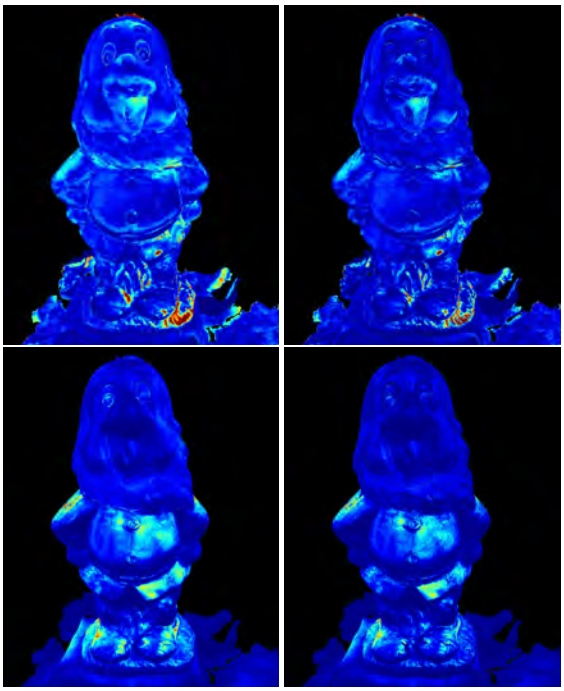


Figure 8: *Change fields without (left) and with the weighting function $w_{diff}(p,q)$. (Top) Diffuse color change field $\delta D(p)$. (Bottom) Residual color change field $\delta HS(p)$.*

we acquired a dataset of photos with its natural appearance and a second one by making some regions more shiny or opaque using oil and matting spray. The regions affected by these changes are highlighted in blue in Figure 9. These regions are the right arm and shoe, which have become more shiny, and the left part of the jacket, trousers and the left shoe, which have become more opaque. The first dataset is composed by 94 photos while the second one by 80 photos. For each dataset, we generated a dense point cloud with Agisoft Photoscan using the input images and a triangular mesh using the dense cloud as input for the Screened Poisson Surface Reconstruction algorithm [KH13]. We created two meshes respectively of 5M triangles for dataset A and 4M triangles for dataset B. The two SLFs were computed in about 5 minutes while the comparison was done in less than one minute using a PC equipped with an Intel(R) Core(TM) i7-4790K CPU (4.00GHz), 32GB RAM, and an NVidia GTX 980 GPU. Figure 9 shows a comparison of the estimated change fields with an input images where the changed regions are highlighted in blue. The change fields give some cues on these regions even if they are not very accurate and precise. Then in some areas there are some false-positive cases due to small differences in the lighting conditions between the two captures that we are not able to normalize with the procedure described in Section 6. For example, under the nose for the diffuse color and on the back of the jacket for the residual reflectance component. Anyway, these results are a good starting point for further investigations.

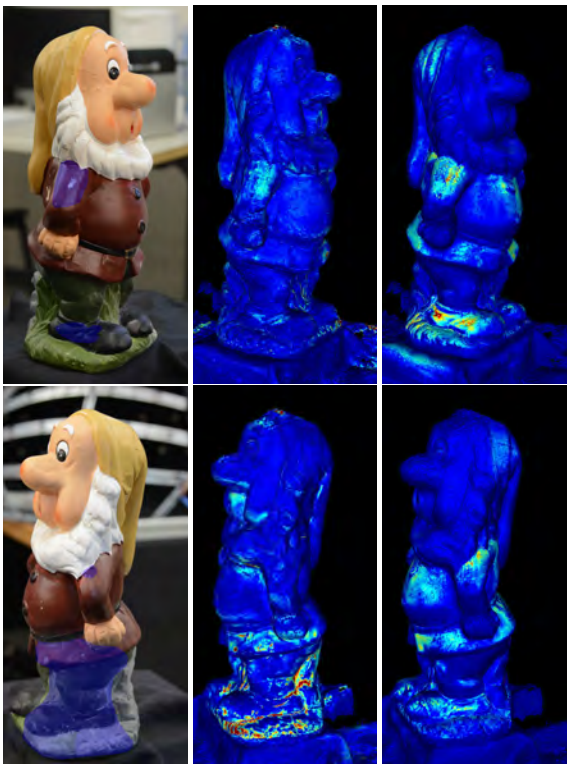


Figure 9: *A comparison of the change fields with the segmentation in the input images of the areas affected by the changes. (Left) Input images with the changed regions highlighted in blue. (Center) Diffuse color change field* $\delta D(p)$*. (Right) Residual color change field* $\delta HS(p)$*.*

8. Conclusion

We have proposed a new method to detect the appearance changes over the time of an object based on the comparison of the Surface Light Fields. Starting from sets of multi-view photos acquired in different time, we compute the 3D model by multi-view 3D reconstruction and the SLF for each time independently. The SLF is estimated with a new method based on a robust separation between the diffuse color and the residual reflectance effects. Then, the SLFs are compared with a weighted approach taking account of small lighting variations and small misalignments in the color-to-geometry projection. We compute two different change fields on the surface of the object: the change field of the diffuse color and the change field of the residual reflectance effects. The computed fields on a dataset with some synthetic changes show several interesting features that can give the user some cues of the areas affected by the change. Even if these result are not very precise and accurate, with some false-positive detections, they can be used as an initial condition for further processing.

There are several research working directions. The first one is to improve the linearization and normalization step on the input images in order to have colors as comparable as possible among the different captures. Another direction is to improve the estimation of the acquisition lighting environment to obtain a more robust computation of the reflectance behavior of the surface, using also a more complex reflectance model as a SVBRDF. Finally, there is the improvement of the comparison procedure which takes account of other factors such as differences in resolution and in focus.

Acknowledgment

The research leading to these results was funded by EU FP7 project ICT FET Harvest4D (http://www.harvest4d.org/, G.A. no. 323567) that we gratefully acknowledge.

References

[AK95] AACH T., KAUP A.: Bayesian algorithms for adaptive change detection in image sequences using markov random fields. *Signal Processing: Image Communication 7*, 2 (1995), 147 – 160.

[BFY00] BLACK M. J., FLEET D. J., YACOOB Y.: Robustly estimating changes in image appearance. *Computer Vision and Image Understanding 78*, 1 (2000), 8 – 31.

[BM92] BESL P. J., MCKAY H. D.: A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence 14*, 2 (Feb 1992), 239–256.

[BP02] BRUZZONE L., PRIETO D. F.: An adaptive semiparametric and context-based approach to unsupervised change detection in multitemporal remote-sensing images. *IEEE Transactions on Image Processing 11*, 4 (Apr 2002), 452–466.

[CCCS08] CALLIERI M., CIGNONI P., CORSINI M., SCOPIGNO R.: Masked photo blending: mapping dense photographic dataset on high-resolution 3d models. *Computer & Graphics 32*, 4 (Aug 2008), 464–473.

[Cli03] CLIFTON C.: Change detection in overhead imagery using neural networks. *Applied Intelligence 18*, 2 (2003), 215–234.

[FTZ*15] FENG W., TIAN F.-P., ZHANG Q., ZHANG N., WAN L., SUN J.: Fine-grained change detection of misaligned scenes with varied illuminations. In *The IEEE International Conference on Computer Vision (ICCV)* (December 2015).

[GKPB04] GAUTRON P., KŘIVÁNEK J., PATTANAIK S. N., BOUATOUCH K.: A novel hemispherical basis for accurate and efficient rendering. In *Rendering Techniques 2004, Eurographics Symposium on Rendering* (June 2004), pp. 321–330.

[HNR84] HSU Y., NAGEL H.-H., REKERS G.: New likelihood test methods for change detection in image sequences. *Computer Vision, Graphics, and Image Processing 26*, 1 (1984), 73 – 106.

[KH13] KAZHDAN M., HOPPE H.: Screened poisson surface reconstruction. *ACM Trans. Graph. 32*, 3 (July 2013), 29:1–29:13.

[KV02] KASETKASEM T., VARSHNEY P. K.: An image change detection algorithm based on markov random field models. *IEEE Transactions on Geoscience and Remote Sensing 40*, 8 (Aug 2002), 1815–1823.

[LKG*03] LENSCH H. P. A., KAUTZ J., GOESELE M., HEIDRICH W., SEIDEL H.-P.: Image-based reconstruction of spatial appearance and geometric detail. *ACM Transactions on Graphics 22*, 2 (Apr. 2003), 234–257.

[LL02] LI L., LEUNG M. K. H.: Integrating intensity and texture differences for robust change detection. *IEEE Transactions on Image Processing 11*, 2 (Feb 2002), 105–112.

[MDS15] MELLADO N., DELLEPIANE M., SCOPIGNO R.: Relative scale estimation and 3d registration of multi-modal geometry using growing least squares. *IEEE Transactions on Visualization and Computer Graphics PP*, 99 (2015), 1–1.

[MN99] MITSUNAGA T., NAYAR S. K.: Radiometric self calibration. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.* (1999), vol. 1, p. 380 Vol. 1.

[PCDS12] PALMA G., CALLIERI M., DELLEPIANE M., SCOPIGNO R.: A Statistical Method for SVBRDF Approximation from Video Sequences in General Lighting Conditions. *Computer Graphics Forum 31*, 4 (2012), 1491–1500.

[PDCS13] PALMA G., DESOGUS N., CIGNONI P., SCOPIGNO R.: Surface light field from video acquired in uncontrolled settings. In *Digital Heritage International Congress (DigitalHeritage), 2013* (Oct 2013), vol. 1, pp. 31–38.

[PGB03] PÉREZ P., GANGNET M., BLAKE A.: Poisson image editing. *ACM Trans. Graph. 22*, 3 (July 2003), 313–318.

[RI03] ROSIN P. L., IOANNIDIS E.: Evaluation of global image thresholding for change detection. *Pattern Recogn. Lett. 24*, 14 (Oct. 2003), 2345–2356.

[Ros02] ROSIN P. L.: Thresholding for change detection. *Computer vision and image understanding 86*, 2 (2002), 79–95.

[RT06] RONG G., TAN T. S.: Jump flooding in GPU with applications to voronoi diagram and distance transform. In *Proceedings of the 2006 Symposium on Interactive 3D Graphics* (2006), Olano M., Séquin C. H., (Eds.), ACM, pp. 109–116.

[SGSC15] STENT S., GHERARDI R., STENGER B., CIPOLLA R.: Detecting change for multi-view, long-term surface inspection. In *Proceedings of the British Machine Vision Conference (BMVC)* (September 2015), BMVA Press, pp. 127.1–127.12.

[SGSC16] STENT S., GHERARDI R., STENGER B., CIPOLLA R.: Precise deterministic change detection for smooth surfaces. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)* (March 2016), pp. 1–9.

[SIN89] SINGH A.: Review article digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing 10*, 6 (1989), 989–1003.

[SO15] SAKURADA K., OKATANI T.: Change detection from a street image pair using CNN features and superpixel segmentation. In *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015* (2015), pp. 61.1–61.12.

[SOD13] SAKURADA K., OKATANI T., DEGUCHI K.: Detecting changes in 3d structure of a scene from multi-view images captured by a vehicle-mounted camera. In *IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, 2013), IEEE, pp. 137–144.

[Str00] STRINGA E.: Morphological change detection algorithms for surveillance applications. In *Proceedings of the British Machine Vision Conference 2000, BMVC 2000, Bristol, UK, 11-14 September 2000* (2000), pp. 1–10.

[SWD05] SHARMA G., WU W., DALAL E. N.: The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application 30*, 1 (2005), 21–30.

[TBP11] TANEJA A., BALLAN L., POLLEFEYS M.: Image based detection of geometric changes in urban environments. In *IEEE International Conference on Computer Vision* (Barcelona, 2011), IEEE, pp. 2336–2343.

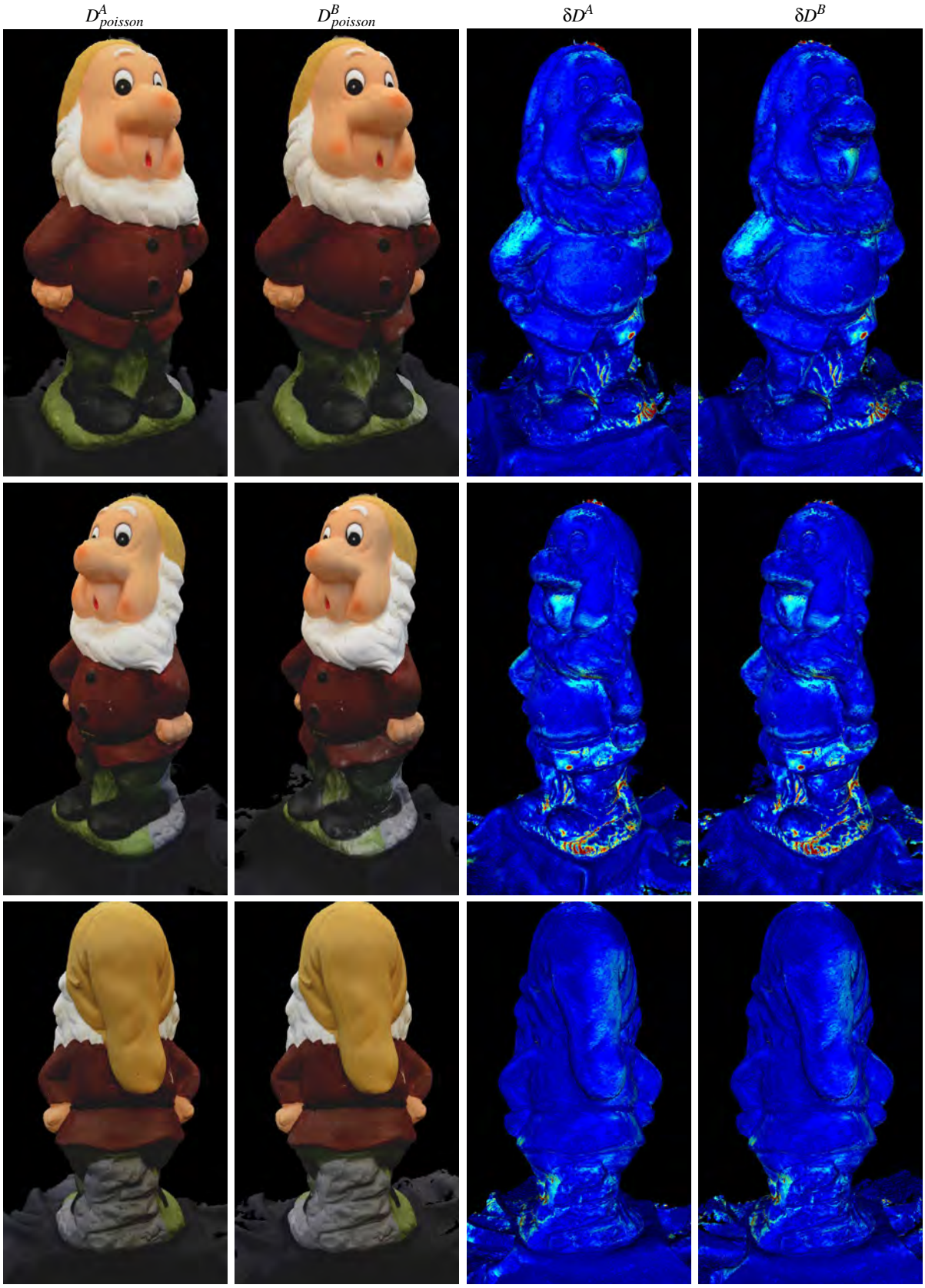


Figure 10: *An example of a change field computed on the diffuse color. The change field is mapped on a color ramp from blue ($\delta D = 0$) to red ($\delta D = 15$). (First column) Rendering of the diffuse color for time step A. (Second column) Rendering of the diffuse color for time step B. (Third column) Change field of the diffuse color of dataset A. (Fourth column) Change field of the diffuse color of dataset B.*

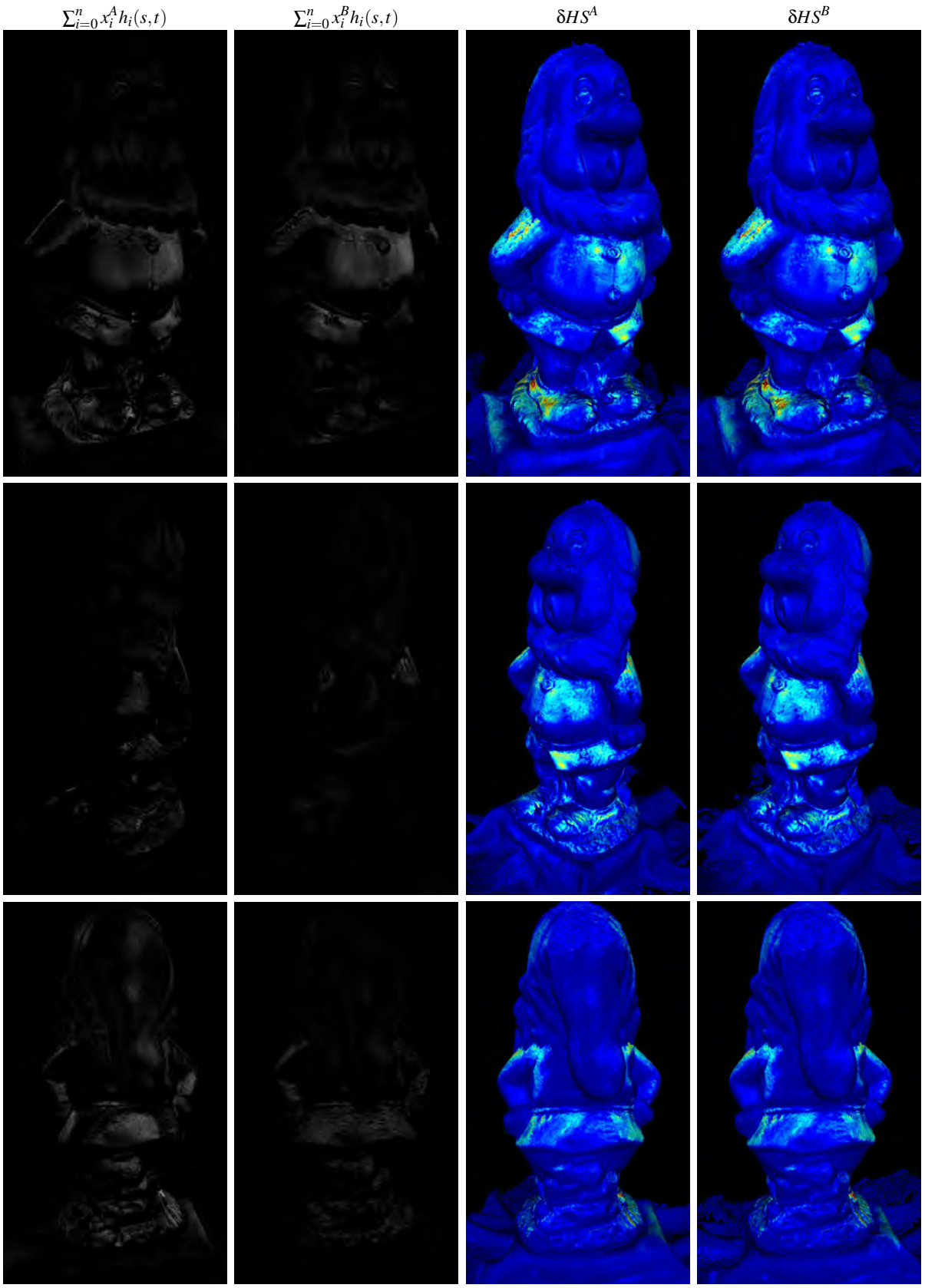


Figure 11: *An example of a change field computed on the residual component of the SLF. The change field is mapped on a color ramp from blue ($\delta HS = 0$) to red ($\delta HS = 0.25$). (First column) Rendering of the residual SLF for time step A. (Second column) Rendering of the residual SLF for time step B. (Third column) Change field of the residual SLF of dataset A. (Fourth column) Change field of the residual SLF of dataset B.*

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], multi-view 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

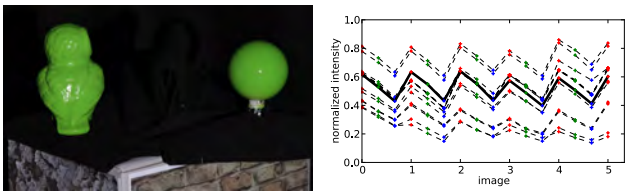


Figure 1. *Left*: Target object and a reference sphere with same reflectance. The high-frequency pattern at the bottom is used to estimate camera pose. *Right*: Some samples from the database of reference profiles (dashed) and a candidate profile (solid).

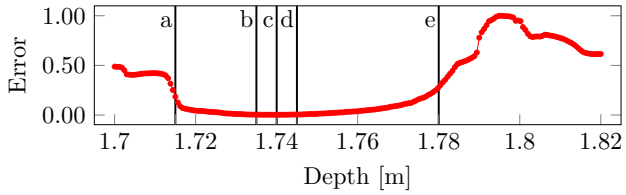


Figure 2. The error of best matching reference profiles along a ray from the camera has a wide basin with very similar error scores. The vertical lines correspond to the depth values in Figure 5.

candidate d we project its 3D position $d \cdot r$ into all m images to obtain intensities $I_j(P_j(dr)), j \in \{1, \ldots, m\}$ 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

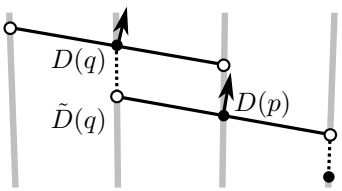


Figure 3. Each ray has a little plane attached. The estimated depth of neighboring pixels should be close to the intersections of their rays with the plane.

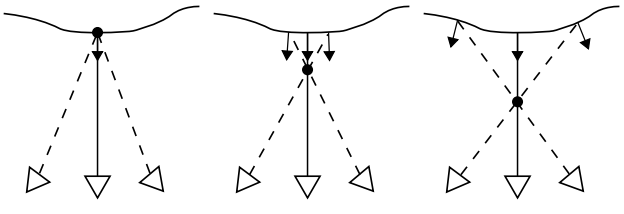


Figure 4. Projections at different depth. (a) All cameras observe the same point. The matching error is zero. (b) Cameras observe different points, but with similar normals. The matching error is still low. (c) Cameras observe points with significantly different normals. The matching error is high.

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

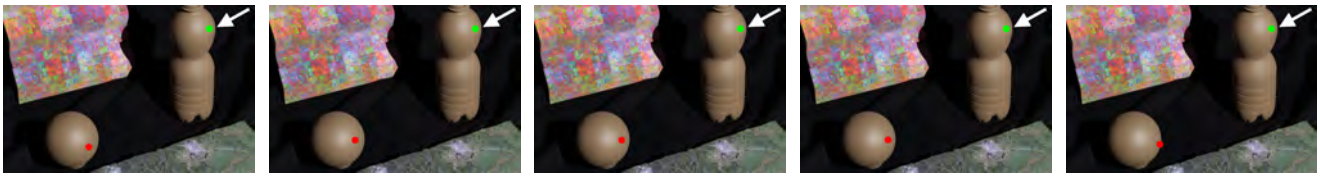


Figure 5. Along the ray going from the camera through the pixel marked in green, the normal corresponding to the best matching reference profile is visualized (red) for increasing depth. Images from left to right correspond to depth a-e in Figure 2. Close to the surface, normals are very stable and similar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

Dataset	Pixels in mask	Energy after iteration 0	10	50	Time [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 $\sim$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.

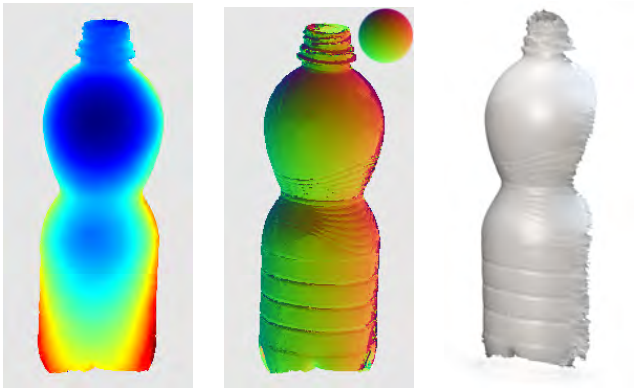


Figure 7. Results for the *bottle* dataset. Left to right: Colored depth map from blue (near) to red (far), the normal map, and a rendering of our triangulated geometry from a novel view.

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.

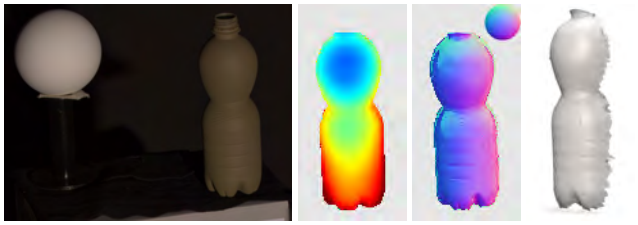


Figure 8. Matching different BRDFs. *Left to right*: An input image showing the diffuse white sphere next to the slightly shiny bottle, the recovered depth map (blue: near, red: far), the normal map, and a rendering from a novel view point.

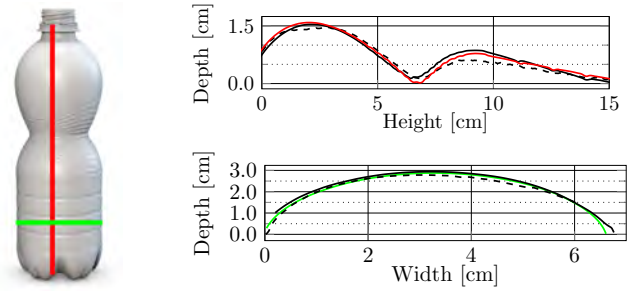
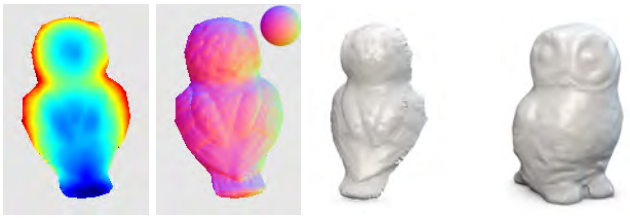


Figure 9. *Left*: Ground truth acquired from structured light scanning with horizontal (green) and vertical (red) profile lines. *Right*: The vertical (top) and horizontal (bottom) cuts through the ground truth (colored) and our depth map (black) show a deviation of less than 2.5 mm for the *bottle*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

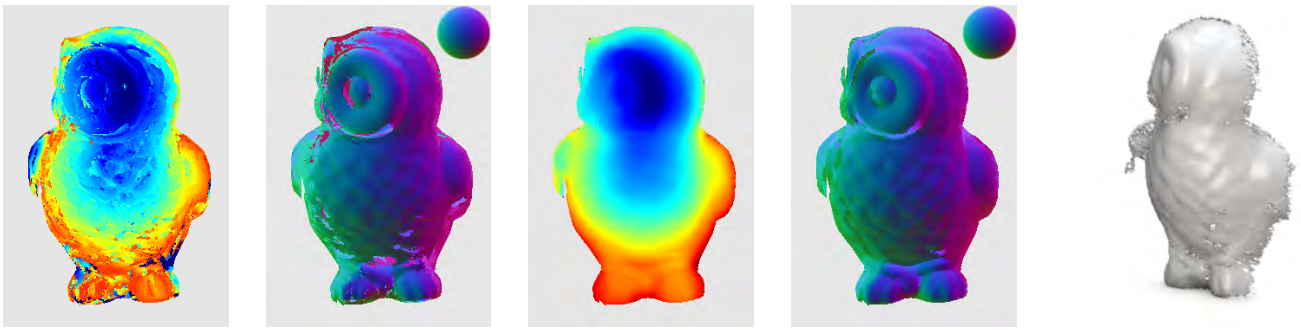


Figure 10. Improvement through optimization. *From left to right*: The initial depth and normal map for the *diffuse owl* dataset; our final depth and normal map after 50 iterations; the triangulated depth map rendered from a novel view.

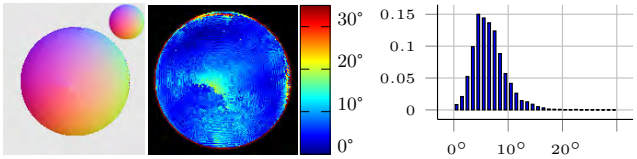
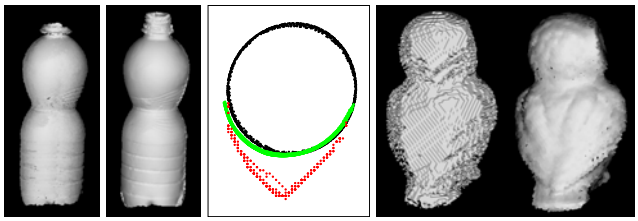


Figure 12. *Left*: Resulting normal map for the *spheres* dataset. *Middle*: The angular error compared to an ideal sphere. *Right*: Histogram over all angular errors below 30° for the *sphere*.

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



(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).

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.

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