Recovering Shape and Spatially-Varying Surface Reflectance under Unknown Illumination

Rui Xia[1,2] Yue Dong[2] Pieter Peers[3] Xin Tong[2]
[1]University of Science and Technology of China [2]Microsoft Research, Beijing [3]College of William & Mary

Figure 1: *Shape and spatially-varying surface reflectance recovered from a video of a rotating object under unknown illumination: (a) a "Bronze Cup" exhibiting both sharp as well as rough specular reflectance, (b) "Ornamental Metal Cup" and (c) a "Carved Disc" with fine geometrical details, (d) a "Cloisonné Bell" with rich texture detail and spatially-varying reflectance properties, and (e) a "Clay Teacup" dominated by diffuse reflectance properties.*

Abstract

We present a novel integrated approach for estimating both spatially-varying surface reflectance and detailed geometry from a video of a rotating object under unknown static illumination. Key to our method is the decoupling of the recovery of normal and surface reflectance from the estimation of surface geometry. We define an apparent normal field with corresponding reflectance for each point (including those not on the object's surface) that best explain the observations. We observe that the object's surface goes through points where the apparent normal field and corresponding reflectance exhibit a high degree of consistency with the observations. However, estimating the apparent normal field requires knowledge of the unknown incident lighting. We therefore formulate the recovery of shape, surface reflectance, and incident lighting, as an iterative process that alternates between estimating shape and lighting, and simultaneously recovers surface reflectance at each step. To recover the shape, we first form an initial surface that passes through locations with consistent apparent temporal traces, followed by a refinement that maximizes the consistency of the surface normals with the underlying apparent normal field. To recover the lighting, we rely on appearance-from-motion using the recovered geometry from the previous step. We demonstrate our integrated framework on a variety of synthetic and real test cases exhibiting a wide variety of materials and shape.

Keywords: Shape-from-shading, Appearance-from-motion, Unknown Lighting

Concepts: •**Computing methodologies** → **Shape inference;** *Reflectance modeling;*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org. © 2016 ACM.
SA '16 Technical Papers, December 05-08, 2016, Macao
ISBN: 978-1-4503-4514-9/16/12

1 Introduction

Digitally reproducing the detailed and intricate appearance of real-world objects has received considerable attention in the past decades. To date, the most successful appearance modeling approaches are measurement-based methods that infer detailed descriptions of the object's surface reflectance from a set of photographs of the object. The majority of measurement-based appearance modeling methods rely on active illumination, precluding appearance acquisition of objects under bright uncontrolled lighting conditions.

Recently, Dong et al. [2014] proposed *appearance-from-motion*, a framework for recovering the spatially-varying appearance from a video of a rotating object under unknown and uncontrolled static illumination. They formulate the recovery of the appearance and lighting as an iterative process that alternates between estimating each component while keeping the other fixed. A key observation is that the effects of lighting discontinuities on the observed temporal changes in the appearance of a surface point are closely related to the form of the point's surface reflectance. However, tracking the temporal changes of a moving surface point requires precisely registered geometry with accurate surface normals. In-situ estimation of shape *a priori*, without making assumptions on the surface reflectance, is difficult due to the complex interplay of shape, reflectance, and lighting in the observations.

In this paper we propose a novel integrated framework for recovering *both* the shape and spatially-varying isotropic surface reflectance from a video of a rotating object under *unknown* static illumination. Inspired by Dong et al. [2014], we propose to estimate both the surface normal and surface reflectance from the temporal trace of a surface point's appearance. However, the position of the surface point is also unknown, without which we are unable to correspond the observations over time, and thus we cannot recover the temporal trace. Instead, we consider the apparent temporal trace for each point in space whether or not it lies on the surface. A key observation is that surface reflectance, normal variation, and position affect the relation between incident lighting and apparent temporal trace differently: surface reflectance smooths incident lighting, nor-

DOI: http://dx.doi.org/10.1145/2980179.2980248

mal variation introduces an offset between the lighting and the apparent temporal trace, and position introduces incoherent discontinuities in the apparent temporal trace when not on the surface. This allows us to construct an apparent normal field with corresponding reflectance functions for each point in the bounding volume that best explain the observed offsets for the temporal traces with respect to the incident lighting. Ideally, the surface shape should go through the points that minimize incoherent discontinuities in the apparent temporal traces while maximizing coherence with the corresponding apparent normals. However, estimating the apparent normal field requires knowledge of the incident lighting, which is also unknown. To resolve this dilemma, we alternate between refining the shape guided by the apparent normal field and estimating the incident lighting using appearance-from-motion [Dong et al. 2014].

The proposed integrated framework for shape and reflectance recovery provides a low-cost solution to measurement-based appearance modeling, only requiring a video camera, without sacrificing quality for a wide range of materials, as showcased by the examples in Figure 1.

2 Related Work

Appearance and shape acquisition methods can roughly be categorized based on whether or not incident illumination is controlled during acquisition. Both classes of acquisition methods are complementary, suited for different applications, and aimed at different operating conditions. The proposed method falls in the category of uncontrolled incident lighting. For sake of brevity, we focus our discussion of prior work on methods that recover shape and/or reflectance under *uncontrolled* lighting – a comprehensive overview of active illumination methods for geometry and appearance acquisition can be found in [Weinmann and Klein 2015].

Shape Estimation under Uncontrolled Lighting A highly successful and popular class of methods for 3D shape recovery under uncontrolled and unknown lighting is multi-view stereo [Seitz et al. 2006]. These methods triangulate the 3D position of surface points by finding corresponding projections in different views. Multi-view stereo methods work best for diffuse-like richly textured objects. Furthermore, while multi-view stereo methods can reconstruct submillimeter accurate 3D shapes, the resulting surface normals can still exhibit significant errors, which makes the resulting geometry ill-suited for reflectance estimation.

In contrast to multi-view stereo, photometric stereo [Woodham 1980] directly estimates surface normals from measurements. Lu et al. [2013] propose a photometric stereo variant for recovering surface normals for unknown isotropic materials from a large number of observations (≈ 150) under uncalibrated (but quasi-uniformly distributed) directional light sources. Basri et al. [2007] propose a photometric stereo solution, limited to Lambertian surface reflectance, from multiple (4 or 9) observations under different (but unknown) lighting conditions using a low-order spherical harmonics representation. Both methods only estimate surface normals for a fixed view. In contrast, we recover the full 3D shape of the subject as well as its spatially-varying surface reflectance.

Example-based photometric stereo [Hertzmann and Seitz 2003; Treuille et al. 2004] reconstructs shape and reflectance of a homogeneous object from photographs of the object under unknown lighting and a reference object with known shape and similar material. Instead of using an explicit reference object, Ackermann et al. [2012] use the portion of the object that can be reliably reconstructed using multi-view stereo as the reference object, and recover the shape of the remainder of the object by integrating the example-based photometric normals. These methods focus on recovering surface normals, and they do not recover lighting and surface reflectance separately.

Reflectance Recovery under Uncontrolled Lighting Both Romeiro et al. [2010] and Lombardi et al. [2016] recover surface reflectance under uncontrolled lighting for general surface reflectance. However, both methods are limited to *homogeneous* materials only. Palma et al. [2012] estimate lighting and clustered surface reflectance from a video of an object with known geometry recorded under unknown uncontrolled lighting. The specular component of the reflectance is modeled by the Phong BRDF. Similarly, Dong et al. [2014] also recover lighting and surface reflectance from a video of a rotating object under unknown lighting, but use a more flexible data-driven microfacet reflectance model and estimate the reflectance parameters for each surface point separately.

Shape and Reflectance under Uncontrolled Lighting Oxholm and Nishino [2012; 2014] recover both shape and surface reflectance using an expectation-maximization framework from a *homogeneous* object under *known* incident natural lighting. In contrast, the proposed method does not require prior knowledge of the incident lighting, and it is not limited to homogeneous material properties.

Barron and Malik [2015] estimate shape, reflectance and illumination from a single image of an object with piece-wise constant Lambertian reflectance under unknown natural illumination. To find the most likely solution, they make a number of prior assumptions on lighting (i.e., log-likelihood), shape (i.e., smoothness, isotropic normal distribution, and silhouette constraints), and reflectance (i.e., piece-wise smooth consisting of a limited palette of natural colors). Wu et al. [2011] use shape-from-shading to refine a coarse geometry obtained from multi-view stereo on a homogeneous Lambertian object under unknown lighting. Wu et al. exploit the low frequency nature of diffuse irradiance, and recover a spherical harmonics representation from shading cues from the coarse multi-view stereo solution. Xu et al. [2014] improve on [Wu et al. 2011] by including visual hull constraints and by solving lighting and shape simultaneously. Valgaerts et al. [2012] extend the work of Wu et al. to dynamic face capture and handle spatially-varying albedo via clustering. All of these methods rely on diffuse cues to refine the shape, and are unlikely to produce good results in the presence of strong specular reflections. A notable exception is shape-from-specular flow [Adato et al. 2010] which reconstructs the geometry of homogeneous specular objects under unknown natural lighting. However, shape-from-specular flow is limited to mirror-like materials only. Our method is suited for objects that contain both diffuse as well as (glossy) specular surface reflectance.

Chandraker [2014] presents a comprehensive theory on shape and reflectance recovery from motion cues of a homogeneous object under a single directional, but unknown, light source. Wang et al. [2016] propose a related theory for recovering shape and spatially-varying material properties using a light field camera (i.e., translational motion) under a *known* single directional light source. The proposed method also relies on motion cues to infer shape and reflectance, but in contrast to Chandraker [2014] and Wang et al. [2016], our method supports more general lighting conditions.

Recently, Wu et al. [2016] recovered shape, normals, lighting, and surface reflectance from RGB-D observations. Due to the reliance of the depth camera on active lighting, their method is only suited for indoor use or under overcast sky. In contrast, the proposed method does not rely on active lighting, and exploits the sparsity of natural lighting in the gradient domain to infer both shape and reflectance.

3 Assumptions

Our goal is to accurately recover the shape and spatially-varying surface reflectance of an object under unknown (and uncontrolled) incident lighting. We desire a solution that is easy to use and applicable in bright uncontrolled environments, precluding the use of active illumination or specialized hardware. In order to keep recovery practical and robust, we make a number of modest assumptions:

Input Data Similar as in [Dong et al. 2014] we take as input a video sequence of an object rotating in front of a static camera. We assume that the camera is calibrated both radiometrically and geometrically. Furthermore, as in [Dong et al. 2014], we also assume that the relative position and rotation of the object with respect to the camera is known for each captured frame. However, unlike Dong et al. we do not require any prior knowledge on the geometry.

Surface Reflectance We assume that the spatially-varying surface reflectance is isotropic and that it can be accurately characterized by a microfacet Bidirectional Reflectance Distribution Function (BRDF) [Nicodemus et al. 1977]:

$$f_r(\omega_i', \omega_o') = \frac{\rho_d}{\pi} + \rho_s \frac{D(\omega_h')G(\omega_i', \omega_o')F(\omega_h')}{4\omega_{i_z}'\omega_{o_z}'}, \qquad (1)$$

for incident and outgoing directions ω_i' and ω_o' – the prime mark indicates that the directions are with respect to the local coordinate frame defined by the surface normal n. ρ_d and ρ_s are the diffuse and specular albedo respectively, D is the microfacet normal distribution function (NDF), G is the shadowing and masking term, and F is the Fresnel reflectance with a fixed index of refraction of 1.3. We follow [Dong et al. 2014] and store the NDF as a 1D tabulated monotonically decreasing function and compute shadowing and masking as in [Ashikhmin et al. 2000]. To model spatially-varying reflectance, we assign separate parameters (n, ρ_d, ρ_s) and an NDF to each surface point p.

Incident Lighting We assume that the incident lighting $E(\omega_i)$ is distant, temporally static, and color-neutral on average. Similar to prior work, we ignore interreflections. Furthermore, we assume that the lighting contains strong discontinuities/edges and that each surface point is *"scanned"* a number of times from different directions by one or more discontinuities.

Global Coordinate Frame We employ an object-relative global coordinate frame, i.e., the position of the surface remains fixed while the (relative) position and orientation of the camera and lighting varies over time t. For brevity, we also mix global and local coordinate frames in expressions, and assume implicit conversion by: $\omega' = R_{n,t}(\omega)$, where $R_{n,t}$ is the product of the transformation R_n to the shading frame defined by the surface normal n, and the transformation R_t due to relative object rotation.

4 Overview

Problem Statement Recovery of the position x, and normal direction n, for each surface point p, and its appearance information (including diffuse albedo ρ_d, specular albedo ρ_s and NDF D), under the unknown natural lighting $E(\omega_i)$ from a video sequence $I(x_p,t)$, can be formulated as an inverse rendering problem:

$$\underset{\{x,n;\rho_d,\rho_s,D\}_p,E(\omega_i)}{\operatorname{argmin}} \sum_p \sum_t ||I(x_p,t) - L(p,t)||^2, \qquad (2)$$

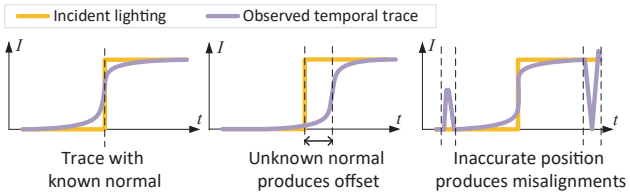

Figure 2: ***Key Observations***. *With the correct normal, the discontinuities in the observed temporal trace and incident lighting are perfectly aligned. Changes in the surface normal direction shifts the observed temporal trace and produces an offset. Inaccurate surface position will produce misalignments which are sparse and cannot be aligned with the lighting.*

where $I(x_p,t)$ is the pixel value observed when back-projecting the position x_p of a point p toward the camera at time t. $L(p,t)$ is the predicted observed radiance at a point p and at time t:

$$L(p,t) = \int_\Omega f_r(\omega_i', \omega_o'(x_p,t); \{\rho_d, \rho_s, D\}_p) E(\omega_i) \omega_{i_z}' d\omega_i'. \qquad (3)$$

The outgoing direction $\omega_o'(x_p,t)$ is defined by the relative camera position at time t and the point's position x_p.

Temporal Trace Directly solving Equation (2) is difficult due to the intricate interplay between shape (x and n), surface reflectance (f_r) and incident lighting (E) in producing the observations I, as well as several ambiguities between the different components. Dong et al. [2014] rely on high frequency discontinuities in the incident lighting to disambiguate lighting and surface reflectance, and show that the temporal changes in the observations are closely related to the NDF:

$$\nabla_t T_{x_p}(t) \approx \rho_s \int_\Omega f_s(\omega_i', \omega_o'(p,t)) \omega_i'_z \nabla_t E(\omega_i) d\omega_i', \qquad (4)$$

where $T_{x_p}(t) = I(x_p,t)$ is the *temporal trace* of the point p over time t, and $f_s(\omega_i', \omega_o')$ is the specular component of the surface reflectance (Equation (1)). Intuitively, the temporal trace is a record of the appearance variations over time of a particular surface point. Hence, given the lighting and geometry, the specular surface reflectance (determined by the NDF) can be computed from the temporal trace. To obtain a robust estimate, Dong et al. employ a shock filter to identify and focus computations on regions of strong gradient.

We also rely on the temporal trace, not only to estimate surface reflectance and lighting, but also to recover shape and surface normals. Our key insight is that surface position and surface normal affect the temporal trace differently (Figure 2). Changes in the surface normal direction offsets the temporal trace with respect to the incident lighting, because the incident lighting $E(\omega_i)$ in Equation (4) is expressed in the global coordinate frame while the BRDF is expressed with respect to the local coordinate frame, which are related by the transformation $R_{n_p,t}$ that depends on the surface normal and the relative camera location at time t. The effects of surface normal (offset) and surface reflectance (blurring) on the temporal trace are essentially orthogonal effects that allow us to robustly estimate both surface normal and surface reflectance from the observed temporal trace of a surface point given the incident lighting.

Accurate knowledge of the surface position is required to assemble the temporal trace of a surface point; discrepancies in the surface position affect the projection onto the observed images, and hence the temporal trace. We define the *apparent temporal trace* $T_x(t)$ of

a 3D point (denoted as x to differentiate from a surface point x_p) as the projection of the 3D point transformed by the relative camera transformation at time t to the corresponding recorded image frame. Clearly, when the point with coordinate x falls on the surface, then the apparent temporal trace is equal to the temporal trace of the corresponding surface point. We observe that the apparent temporal trace for off-surface points exhibits additional discontinuities incoherent with the incident lighting due misaligned projections that cross material boundaries (Figure 2, right). These discontinuities can typically not be explained via a combination of normal offsets and surface reflectance functions. The residual error induced by these unexplained discontinuities with respect to incident lighting, can thus serve as an indicator whether a point lies on the surface.

Algorithm Overview The above observations suggest that we can estimate an apparent normal and apparent surface reflectance for each point in space, and use the residual error as a guide to find the object's surface. However, this requires knowledge of the incident lighting. Appearance-from-motion [Dong et al. 2014] has shown that, given known geometry, we can estimate the incident lighting and surface reflectance from the temporal traces. Since the recovery of shape and lighting depend on the other, we formulate the recovery as an iterative process, that alternates between both:

1. Finding the most likely temporal traces (and thus geometry and normals), given the estimated lighting, from all possible apparent temporal traces (Section 5):

$$\underset{\{x,n\}_p}{\operatorname{argmin}} \sum_p \sum_t ||T_{x_p}(t) - L(p,t)||^2, \tag{5}$$

   with known incident lighting $E(\omega_i)$; and,
2. Determining the most likely incident lighting and surface reflectance, given the object's geometry and temporal traces (Section 6):

$$\underset{\{\rho_d,\rho_s,D\}_p,E}{\operatorname{argmin}} \sum_p \sum_t ||I(x_p,t) - L(p,t)||^2, \tag{6}$$

   with known geometry: $\{x,n\}_p$.

Initialization To bootstrap the alternating optimization process, we provide initial estimates of both geometry and lighting. We initialize the geometry with the visual hull computed from silhouettes obtained by background subtraction. We initialize the lighting using a single iteration of the bootstrapping process from [Dong et al. 2014] using the initial visual hull geometry to establish the temporal traces. We briefly summarize the steps below, and refer to [Dong et al. 2014] for a more detailed description:

1. A per surface point local incident lighting estimate is established by first projecting the temporal trace to the sphere of directions (based on relative camera orientation), then widening the trace by copying the nearest value from the projection, and finally, shock filtering the expanded projection to enhance discontinuities.
2. Each local lighting estimate is scaled by an unknown and potentially different albedo. We remove this effect (up to an unknown scale factor), by performing a global optimization to scale each local incident lighting estimate such that differences on overlapping sections are minimized.
3. Next, we search for the *unnormalized* NDF $\bar{D}$ that best explains the observations, given the local incident lighting estimate (Equation (4)).
4. Given the unnormalized NDFs, we then recover the (global) incident lighting $E(\omega_i)$ using a multi-resolution deconvolution-based minimization:

$$\underset{E(\omega_i)}{\operatorname{argmin}} \sum_t \sum_x w(x_p)||I(x_p,t) - L_s(p,t)||^2 + \lambda||\nabla_\omega E(\omega_i)||^{0.8}, \tag{7}$$

   where $L_s(p,t)$ is the outgoing radiance from the specular component of the BRDF determined by the unnormalized NDF $\bar{D}$. We consider points which exhibit a sharp specular BRDF and which are "scanned" by strong edges in the lighting to be more reliable: $w(x_p) = \frac{1}{\sigma}\sum_t \nabla_t T_{x_p}(t)$, with σ the variance of the NDF. As in Dong et al., the number of multi-resolution scales (and thus resolution of the recovered lighting) is determined by the bandwidth of the estimated NDFs to avoid overfitting.
5. Finally, we subtract the maximal diffuse lighting to compensate for the baked-in diffuse reflectance in the local incident lighting estimates.

5 Shape and Normal Reconstruction

As observed in Section 4, shape and reflectance influence the apparent temporal trace differently. A key observation is that error on the position of a surface point introduce discontinuities inconsistent with the incident lighting – changes in surface normal or appearance parameters are unlikely to characterize these discontinuities. This suggest that we can formulate the recovery of shape and surface normals as a two-step process. In the first step we compute the apparent normal field as well as their corresponding apparent reflectance parameters for every point in the bounding volume. In the next step, we find the closed surface through the apparent normal field for which the temporal traces are most consistent with the incident lighting.

5.1 Apparent Normal and Reflectance Field

To compute the apparent normal and reflectance field, we first discretize the maximal bounding volume in a voxel grid. We use the bounding volume of the initial geometry, increased by 10%, as the maximal bounding volume, and discretize it in a 128^3 voxel grid. For each voxel center, we then compute the apparent NDF, albedo, and surface normal.

Apparent NDF We estimate the apparent NDF similarly as in the initialization of the lighting. We construct a local incident lighting approximation directly from the apparent temporal trace: (1) project the apparent trace on the sphere of directions, (2) expand the projection perpendicular to the trace direction, and (3) apply a shock filter to localize discontinuities. We then find the apparent NDF that best explains the observations using Equation (4). However, Equation (4) requires knowledge of the surface normal which is unknown. Fortunately, for the purpose of recovering the apparent surface normal, the exact shape of the NDF is not required as long as the angular extend of the BRDF matches well. By approximating the *unnormalized* apparent NDF as: $\bar{D}(\omega_h{}') \approx f_s(\omega_i{}', \omega_o{}')\omega_{i_z}'$ (i.e., the NDF under normal incidence), it follows that an approximative NDF can be found by normalizing the zero-mean kernel that best relates the derivative of local incident lighting to the derivative of the temporal trace.

Apparent Normal and Albedo Estimation Given the estimated apparent NDF, we then find the optimal apparent diffuse albedo ρ_d and specular albedo ρ_s and the apparent normal direction n_x. However, the effect of the surface normal on the outgoing radiance is highly non-linear. We therefore brute-force search, from a discrete set of candidate normals $\mathcal{N}$ uniformly distributed over the full sphere of directions, for the apparent normal n_x that minimizes

the error $\varepsilon^2_{n_x}$ when estimating the albedo:

$$\varepsilon^2_{n_x} = \underset{n_x \in \mathcal{N}}{\operatorname{argmin}} \left(\underset{\rho_d, \rho_s}{\operatorname{argmin}} \sum_t \frac{1}{N} ||T_x(t) - \rho_d L_d(x,t) - \rho_s L_s(x,t)||^2 \right), \tag{8}$$

where L_d and L_s are the predicted diffuse and specular outgoing radiance at x toward the camera at time t given the NDF. The normalization factor $N = \sum_t T_x(t)$ ensures that $\varepsilon^2_{n_x}$ at different points x are comparable (Section 5.2).

Visibility During optimization we only include frames in the apparent temporal trace for which $(\omega_o(x,t) \cdot n_x) \geq 0$ to avoid including reflectance from back-facing surfaces. However, there still exists an ambiguity in the sign of the normal at each surface point; generally we can find two opposite facing normals with corresponding surface reflectances that match the observations well. We resolve this ambiguity guided by the visibility of the closest projection of the voxel, along the view vector, to the geometry from the previous iteration; we include the frame in the apparent temporal trace if the closest projection is visible from the respective camera.

5.2 Geometry Reconstruction

To reconstruct the geometry, we search for the closed surface based on the the apparent normal field and corresponding error $\varepsilon^2_{n_x}$. Figure 3 shows a 2D slice from three different apparent normal fields and the corresponding errors, as well as the ground truth surfaces with the *Wallpaper* SVBRDF [Dong et al. 2010]. As expected, the ground truth surface passes through regions of low error $\varepsilon^2_{n_x}$. However, the error field can also vary rapidly with position as it is the integration of the (high frequency) discontinuities caused by misalignment. Consequently, the error field is sensitive to measurement noise and the exact location (i.e, voxel center) at which it is computed. Furthermore, by itself the error field is not sufficient to disambiguate the exact location of the surface in regions of little appearance variation or in regions with large errors due to occlusions. Hence, we cannot solely rely on the error field to recover the surface. Whereas the error field is dominated by the influence of sparse high frequency discontinuities, the normal field in contrast, is mainly influenced by difference in offset over the (whole extent of the) apparent temporal trace and the BRDF filtered incident lighting. Consequently, the apparent normal field is more robust to noise and misalignments and varies more smoothly than the error field. We take both the error field and the apparent normal field in account, by formulating the recovery of the geometry as an iterative process where a rough initial shape is constructed based on the error field, and subsequently refined guided by the smooth apparent normal field.

Robust Shape Initialization We construct a coarse initial shape using Poisson surface reconstruction [Kazhdan and Hoppe 2013] on the voxel center positions x with corresponding normal n_x which have an error $\varepsilon^2_{n_x}$ less than a user-specified threshold. We found that a threshold of 2.5×10^{-4} works well in practice. The magnitude of the threshold determines the *"width"* of the cloud of voxels around the ground truth surface. A too small threshold can result in holes that the Poisson reconstruction needs to fill in, while a too large threshold results in a wider cloud, and thus gives more freedom to the Poisson reconstruction to place the surface away from the ground truth surface.

Initializing the shape reconstruction from a "fresh" geometry instead of reusing the geometry from the previous iteration serves two goals. First, it ensures that we start from a geometry that lies in a region with low error $\varepsilon^2_{n_x}$ no matter the quality of the outcome of the previous iteration, or the quality of the initial geometry (i.e., a visual hull in our case). Second, it assures that the vertices are uniformly distributed on the initial mesh, and it avoids clustering of vertices due to repeated refinement.

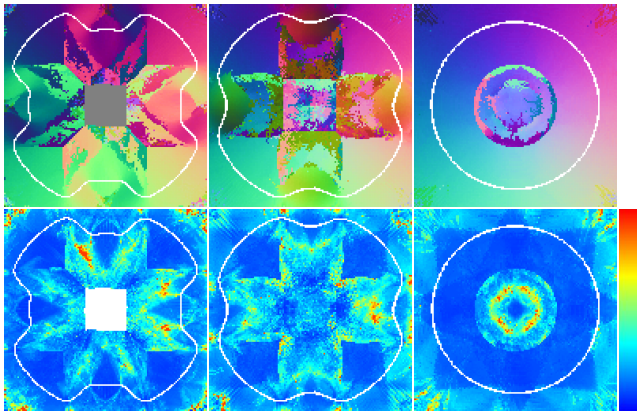

Figure 3: ***Apparent Normal Field and Corresponding Error*** *A slice through the 3D apparent normal field (top), and their corresponding $\varepsilon^2_{n_x}$ (bottom) for three different synthetic objects with the "Wallpaper" SVBRDF.*

Surface Optimization Inspired by [Nehab et al. 2005], we *iteratively* refine the geometry by updating the vertex positions while minimizing the deviation between the surface normal and the underlying apparent normal. Because the apparent normal field varies with position, we also limit vertex-displacement per iteration (effectively linearizing the normal field). We further regularize the optimization by penalizing changes in topology. Formally, for a vertex p, we find the position x_p that minimizes the weighted sum of three constraints:

$$\underset{x_p}{\operatorname{argmin}} \lambda_N \hat{E}_N + \lambda_P \hat{E}_P + \lambda_L \hat{E}_L, \tag{9}$$

where λ_N, λ_P, and λ_L are weights to balance the three constraints – in our implementation we set $\lambda_N = 0.9$, $\lambda_P = 0.1$, and $\lambda_L = 0.2$. We compute the new vertex position using a linear least squares minimization, and formulate the three constraint appropriately:

Normal Constraint $\hat{E}_N$: penalizes deviations from the apparent surface normal. To fit this inherently non-linear constraint in a linear least squares form, we employ the linearized surface normal approximation from [Nehab et al. 2005]:

$$\hat{E}_N = \sum_p \sum_{q,r \in \mathbb{N}(p)} ||n_p^{prev} \cdot (x_q - x_r)||^2, \tag{10}$$

where $\mathbb{N}(p)$ is the one-ring neighborhood of vertices around p. We only compute this constraint for vertices that reside in the set of voxels that are used for constructing the initial shape; we consider voxels outside this selection to have unreliable apparent surface normal estimates. Note that we express this constraint in terms of the apparent normal of the original vertex position (at the beginning of the iteration). Otherwise, the apparent normal would depend on the optimized position x_p, and Equation (10) would be non-linear – this is a reasonable assumption for small changes in vertex position and the overall smooth behavior of the apparent normal field.

Position Constraint $\hat{E}_P$ serves to (1) limit the magnitude in changes to the vertex position such that the apparent normal remains approximately constant (see above), and (2) maintain the overall size of the object – the linearized approximation of the surface normals

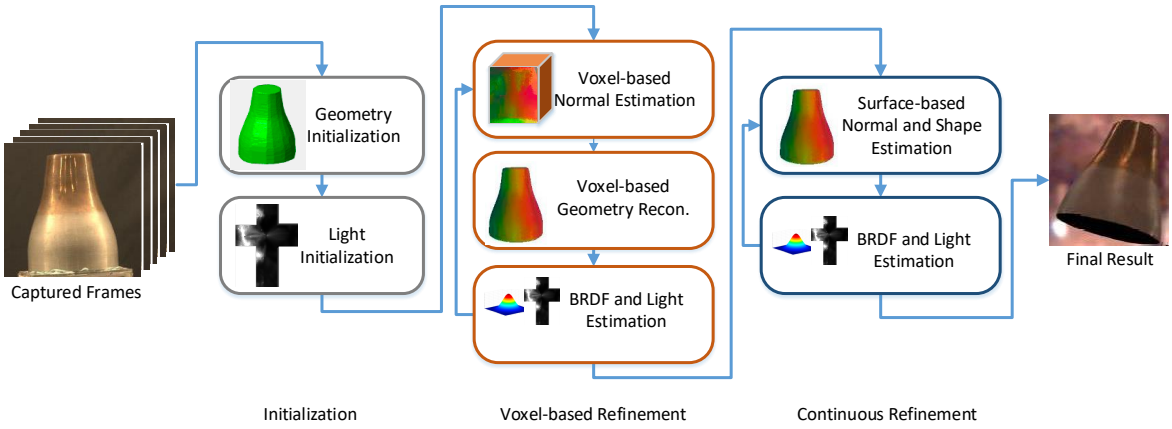

Figure 4: ***Algorithm Overview*** *After initializing the shape (visual hull) and incident lighting, our method iterates between optimizing shape and lighting. Initially, we recover shape using a discrete voxel-based approach. However, once no further improvements can be attained, we switch to a continuous shape refinement. Finally, after the shape has converged, we iteratively refine the incident lighting and appearance until convergence.*

tends to shrink the geometry. The position constraint is defined as the distance between the position of the vertex and its original position:

$$\hat{E}_P = \sum_p ||x_p - x_p^{prev}||^2. \tag{11}$$

Laplacian Constraint $\hat{E}_L$ retains the topology of the geometry and avoids flipped or degraded triangles:

$$\hat{E}_L = \sum_p ||\mathscr{L}(x_p^{prev}) - \mathscr{L}(x_p)||^2, \tag{12}$$

where $\mathscr{L}(\cdot)$ is the Laplacian of the vertex p.

We track convergence by comparing the average distance between the original and updated vertices. To negate the effects of vertices sliding over the surface, we compute the distance along the normal direction: $D(x_p^{prev}, x_p) = ||(x_p - x_p^{prev}) \cdot n_p||^2$. In our implementation we stop refinement when the average distance is less than 10^{-4} times the object size.

Invalid Normal Exclusion The above algorithm relies on the observation that an offset in the discontinuities between the apparent traces and the incident lighting can be explained by normal variations, and conversely, incoherent discontinuities are related to mismatches in the position. However, this observation is only strictly valid for unimodal surface reflectance, while in practice we employ a dichromatic model (i.e., with a diffuse and specular component). In the case where the specular component is weak compared to the diffuse component, it is possible that an incoherent discontinuity results in a larger error than the coherent discontinuities, and thus a relatively smaller overall error can be obtained by incorrectly fitting a stronger "ghost" specular BRDF (with incorrect normal) to the incoherent discontinuity. Such a situation can occur at the interface between diffuse materials with vastly different albedos – any error on the position can result in an apparent trace that crosses the material boundary (i.e., incoherent discontinuity). We therefore exclude any such potentially invalid normal in the above surface optimization. Practically, if the standard deviation in a region around a surface point over the diffuse albedo is above 0.01, we then ignore the normal in the optimization of the corresponding point p by setting the normal weight λ_N to 0, and instead constrain (smooth) the surface locally by increasing the Laplacian weight λ_L to 10. The region size should be set equal or larger than expected surface error – a larger region results in some loss of detail. In our implementation we err on the side of caution and use a fixed conservative region with a radius of 3% of the object size.

6 Appearance and Lighting Estimation

Given the geometry, and thus temporal traces, we refine the estimated lighting and appearance using appearance-from-motion [Dong et al. 2014]. Note that in this step we also re-estimate the appearance for the surface points which do not necessarily lie at the voxel centers (and thus we cannot reuse the estimated apparent reflectance from Section 5.1):

1. Given the current estimate of the (shock filtered) lighting and current set of temporal traces, we compute the unnormalized NDF using Equation (4).
2. Next, we compute the diffuse and specular albedo using Equation (8) using the surface normal from the estimated geometry.
3. Finally, we obtain a new estimate of the lighting by minimizing Equation (7).

Unlike Dong et al. [2014], we only run a single iteration of the above procedure, because the geometry is only an estimate and likely inaccurate, especially for early iterations.

7 Continuous Surface Refinement

We alternate between estimating shape/normals and lighting/appearance until the recovered geometry converges. However, the accuracy of the converged geometry is limited by the voxel grid resolution. In theory we can gradually increase the grid resolution, however, in practice memory and computation requirements become prohibitive for very dense voxel grids. Instead, of gradually increasing the grid resolution, we switch to a continuous geometry estimation after no further improvements can be obtained with the fixed resolution voxel-based geometry estimation (typically in a small number of iterations), and continue to alternate between estimating the shape and lighting/appearance.

We assume that the converged voxel-geometry is close to the ground truth, and therefore directly refine the current geometry instead of generating a new initial shape every iteration. The continuous refinement is identical to the voxel-based refinement algorithm (Section 5.2), except that n_p^{prev} (Equation (10)) is now evaluated on

Figure 5: ***Shape and Reflectance Evolution.*** *Image sequence of intermediate recovered shapes (top) and surface reflectances (bottom) for each iteration starting from the initial visual hull, followed by the voxel-based refinement results, and subsequently, the results from continuous refinement.*

the fly (using the same process as in Section 5.1) instead of being precomputed on grid locations. We store the intermediate spatially-varying surface reflectance estimates in a 256×256 texture.

Finally, once the final geometry has converged (i.e., the average distance is less than 10^{-5}), we rerun the lighting and appearance estimation at quadruple resolution (i.e., in a 1024×1024 appearance texture), without further refining the geometry, until convergence. Figure 4 summarizes our integrated shape and appearance pipeline. Figure 5 shows the evolution of the shape and reflectance recovery over several iterations, until convergence. Notice how large scale features are quickly resolved in just a few iterations, while fine-scale details (e.g., the decorative banding at the top) requires more iterations. In general, the voxel-based refinement quickly converges to a coarse-scale accurate geometry, while the continuous refinement adds the fine-scale details.

8 Implementation Details

By far the computationally most expensive step in our integrated shape and appearance recovery algorithm is the computation of the apparent normals (Section 5.1). We accelerate this computation in two ways: reducing the number of voxels for which we compute the apparent normal field, and by precomputing the convolution of the BRDF with the lighting.

Voxel Reduction With exception of the first iteration, we expect that the robust shape initialization in Section 5.1 will produce a mesh close to the geometry of the previous iteration. We therefore only consider the voxels within 10 units from the previous iteration's geometry for the shape initialization and refinement. Because we do not have an accuracy-guarantee for the initial mesh, we still consider all voxels in the maximal bounding volume for the first iteration.

Precomputation We exploit that the incident lighting is distant, and thus the same in the computation of each apparent normal. Furthermore, we observe that the exact shape of the apparent NDF is less important for the estimation of the apparent normal field. These two observations enable us to significantly speed-up the apparent normal estimation by precomputing the convolution of the incident lighting with a set of basis BRDFs, and express the outgoing radiance due to the apparent BRDF as a linear combination of the precomputed convolutions:

$$L(n) = \rho_d L_d(n) + \rho_s \sum_i w_i L_i(n), \tag{13}$$

where L_i is the outgoing radiance for the i-th basis BRDF with surface normal n, and where the outgoing direction aligns with the optical axis $\bar{\omega}_o$ of the camera:

$$L_i(n) = \int_\Omega f_{ri}(R_n(\omega_i), R_n(\bar{\omega}_o)) E(\omega_i)(\omega_i \cdot n) d\omega_i, \tag{14}$$

where R_n is the transformation into the local frame around n, and where the basis BRDFs are defined by a Beckmann NDF [Cook and Torrance 1982] with a 15 different roughness values distributed logarithmically in the range $[0.01, 0.2]$.

For each point's estimated apparent NDF D_x, we compute the optimal weights w_i, and approximate the outgoing radiance for a posited apparent normal n using the precomputed radiance (Equation (13)), taking in account the relative object rotation at time t. In addition, we correct for differences in the specular reflectance when the outgoing direction ω_o does not align with the precomputed outgoing direction $\bar{\omega}_o$ by using a transformed normal $\hat{n}(\omega_o, n) = n - \frac{1}{2}(\bar{\omega}_o - \omega_o)$ when looking up the precomputed outgoing specular reflectance.

Luminance We further speed up the computation of appearance by exploiting the fact that the reflectances of the different color channels are highly correlated. We therefore, first recover shape and monochrome reflectance from the luminance of the observations. Finally, we perform a single full 3 color channel continuous geometry and reflectance optimization step.

9 Results

Experimental Setup Our method takes as input a video of a rotating object, and the relative pose of the object for each captured frame. We record the video sequence with a Canon EOS 5D Mark II equipped with an EF-100 F2.8 lens, and record single-exposure radiometrically linear RAW images – we manually set the exposure to ensure no pixels are oversaturated during capture. The intrinsic camera parameters are calibrated with the method of Zhang et al. [2000].

To avoid bias in our results due to inaccurate pose calibration, we control the rotation with a Direction Perception Pan-Tilt Unit-D46, and track the relative rotation and position using a hexagonal calibration target. Please refer to the supplementary material for an example of an input video sequence. Each side of the calibration target is printed with a 7×7 grid of rings and a unique ARTag [Fiala 2005]. We use the ARTag to identify which side is visible, and the use the detected ring centers to estimate the relative camera pose using a variant of bundle adjustment [Triggs et al. 1999] – the 3D

location of the ring centers are known, and kept fixed during adjustment. The resulting average calibration error is less than 0.3 pixels, and the maximum error is less than 1 pixel. Note however, that while we employ a controlled-motion setup and a calibration target, our method is independent of the employed pose calibration technique and any sufficiently accurate pose estimation method can be used instead; joint estimation of pose and appearance/shape would be an interesting avenue for future work.

We capture a video sequence of 1243 frames, and ensure that each surface point is "scanned" in at least two directions by edges in the lighting. Processing took approximately 10 hours per dataset on a 20 node PC cluster of dual Xeon E2630 CPUs with 64GB of memory: 2 hours where spend on data preprocessing, 30 minutes on initialization, 3 hours on shape and normal reconstruction, and 4 hours on the continuous surface refinement. We perform, on average, 2 iterations over the voxel grid (Section 5), 6 iterations for the continuous refinement (Section 7), followed by an additional single continuous refinement on all three color channels. The iterative geometry refinement guided by the apparent normal field (for both voxel and continuous) is run for 10 iterations.

Experimental Results We demonstrate our technique on the following objects:

- **Bronze Cup:** exhibiting rich BRDF variations, changing gradually from sharp specular at the top to rough specular at the bottom. Both regions contain other detailed spatially-varying reflectance properties: the sharp specular region contains some less reflective oxidation spots, whereas the rough specular regions are marked by a horizontal stripe pattern.
- **Ornamental Metal Cup:** with fine geometrical details and rough specular surface reflectance.
- **Carved Disc:** exhibiting rich geometrical details that are difficult to recover with silhouette-based or correspondence-based methods.
- **Cloisonné Bell:** with rich texture details and spatially-varying surface reflectance.
- **Clay Teacup:** dominated by diffuse surface reflectance and a weak rough specular component.

The shape and appearance of the *Bronze Cup*, *Ornamental Metal Cup*, *Carved Disc*, and the *Cloisonné Bell* are estimated from a video sequence captured in a windowless hallway illuminated by a long tube light overhead and a small wall light. The video sequence for the *Clay Teacup* was captured in a typical office environment dominated by a large window.

Figure 1 show-cases the recovered geometry and appearance rendered for each of the acquired objects under the *St. Peter's Basilica* light probe. Please refer to the supplemental video for visualizations with dynamic viewpoint and lighting conditions. Figure 6 shows visualizations of each of the components separately (i.e., ground truth and recovered incident lighting, normal map, diffuse and specular albedo, and specular roughness (computed as the standard deviation of the recovered NDF). Note that the recovered normal map includes fine-scale geometrical details, and that the appearance properties vary smoothly, demonstrating the robustness of our method. While the accuracy of the recovered lighting is inherently limited by the bandpass behavior of the BRDFs, the recovered incident lighting for each example shows that the overall placement and size of the recovered lighting features are correctly estimated. Note that for the *Carved Disc* example, the majority of the observed reflections originate from a small cone of lighting directions, and thus the recovered lighting is most accurate for this region, and larger reconstruction error can be observed outside this region. We qualitatively evaluate the recovered shape and surface reflectance by comparing a reference photograph and a rendering under a novel lighting environment (Figure 7).

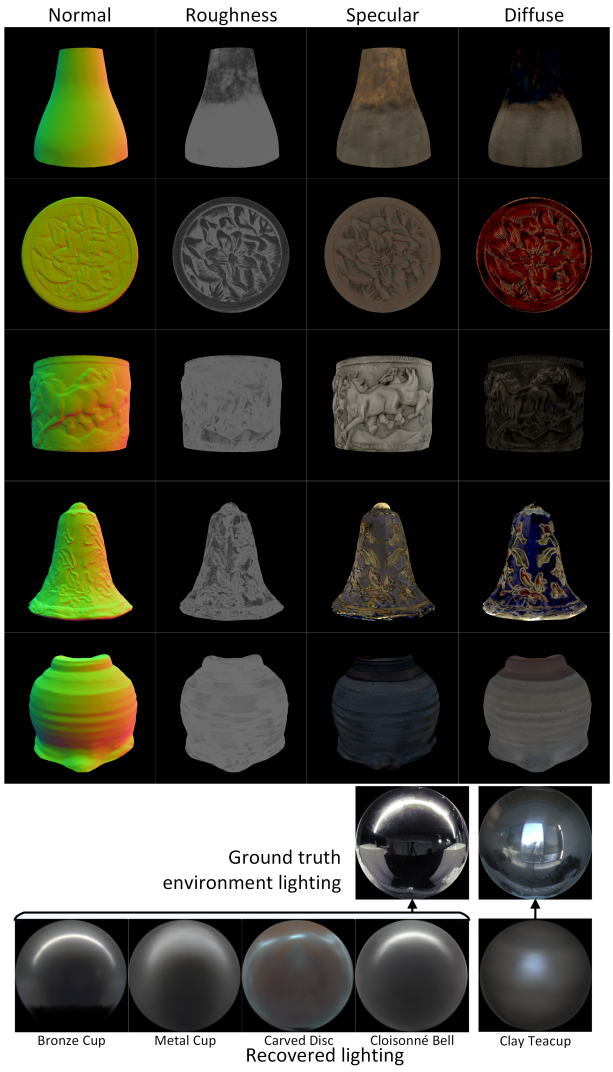

Figure 6: ***Recovered Lighting, Surface Normals, and Reflectance Properties***. *Incident lighting, surface normals, and reflectance properties (diffuse and specular albedo, and specular roughness (computed as the standard deviation of the NDF)) visualized for each of the acquired objects.*

Figure 7: ***Qualitative Validation***. *Side-by-side comparisons of reference photographs and recovered shape and surface reflectance visualized under a novel lighting conditions different from the acquisition environment.*

Table 1: ***Quantitative Error Summary:*** *Shape accuracy is computed as the relative Hausdorff distance (normalized by the length of the largest bounding box axis) for the mean, maximum, and RMS error on the vertex positions. Similarly, surface normal accuracy is computed by the Hausdorff distance on the median normal difference in degrees. Finally, the accuracy of the appearance is expressed as the RMS error on the renderings shown in the respective figures.*

	Position			Normal	Appearance
Result	Mean	Maximum	RMSE	Median	RMSE
Figure 8.a	0.000601	0.005167	0.001043	0.245493	0.0616
Figure 8.b	0.001052	0.005644	0.001322	0.257171	0.0801
Figure 8.c	0.000975	0.007566	0.001145	0.37115	0.0920
Figure 10.a	0.001122	0.006458	0.001526	0.720365	0.0554
Figure 10.b	0.000788	0.005716	0.001366	0.247082	0.0923
Figure 10.c	0.000568	0.004855	0.001004	0.245493	0.0336
Figure 10.d	0.002454	0.006298	0.002924	0.736482	0.0250
Figure 9.b	0.001492	0.007194	0.001858	0.298051	0.1191
Figure 9.d	0.000870	0.005241	0.001216	0.396636	0.1375
Figure 9.e	0.001364	0.005138	0.001668	0.294751	0.1221
Figure 12.a	0.004249	0.014307	0.005013	0.244694	0.1411
Figure 12.b	0.004039	0.014441	0.004945	0.243893	0.1522

10 Discussion

The experimental results demonstrate the quality and potential of our shape and appearance reconstruction method on a variety of objects. To better quantify the robustness and accuracy of our method, we perform an in-depth study on synthetic datasets that allow us to study each component in isolation. In particular we explore: the accuracy of the geometry and appearance reconstruction, and the robustness to the pose calibration, lighting, initialization, and rotation speed.

We employ the following error metrics:

- **Shape:** To avoid bias due to misalignment and/or differences in mesh-resolution, we will use the Hausdorff distance with an Euclidian distance metric and a local search range of 1% on the recovered object normalized by the length of the longest axis of the bounding box.
- **Normal:** Similar as for the shape error, we employ the Hausdorff distance (i.e., 1% normalized by bounding box) and use the angle difference between normals as a distance metric.
- **Appearance:** Instead of directly measuring the distance between two reflectance functions, which is an open research problem, we quantize the difference in appearance by computing the RMS error on renderings of the recovered and ground truth object.

10.1 Accuracy Validation

Impact of Shape We validate the impact of the object's shape on the accuracy of the recovered geometry and appearance by computing the reconstruction error on three different shapes with different types of geometrical features: two spherical shapes with (gradual and sharp) concavities, and an object with sharp and rounded corners. In all three cases we use the Wallpaper SVBRDF dataset [Dong et al. 2010] which contains a mixture of diffuse and highly specular features, and simulate a video sequence illuminated by the "Uffizi Gallery" light probe. Figure 8 shows visualizations of the ground truth and recovered geometry for the three objects under novel lighting. Overall, the recovered shape and surface reflectance closely matches that of the ground truth. However, visually, we can detect larger error around the sharp edges in Figure 8.c. Sharp geometrical features like these require either careful placement of the vertices to align them with the sharp features, or a very high mesh density. A similar effect can be detected on the *Carved Disc* in Figure 1. Table 1, rows 1-3, summarize the error statistics for the different recovered shapes and surface reflectances. For comparison, Figure 9.c shows the result of directly applying appearance-from-

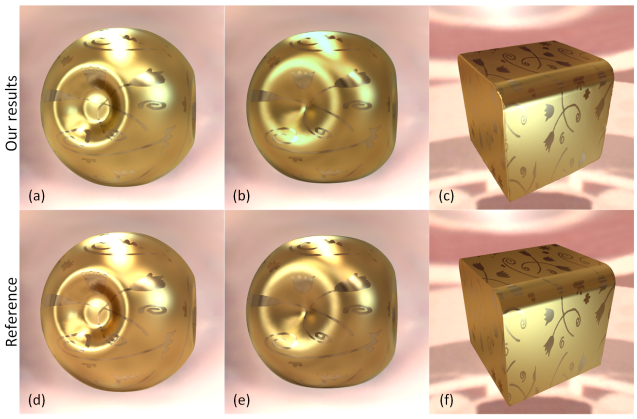

Figure 8: ***Impact of Object Shape****. Visualizations of recovered shape and appearance for three different object shapes exhibiting different geometrical characteristics compared to ground truth visualizations.*

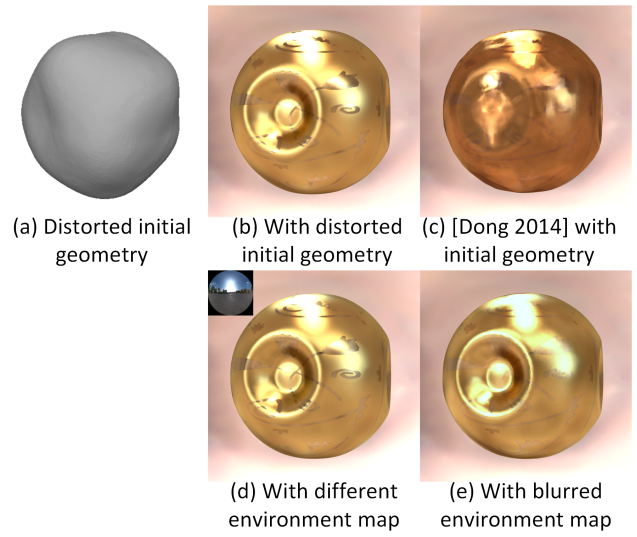

Figure 9: ***Robustness to Input****. Our method is robust to differences in the initial geometry and the lighting environment under which the video sequence is captured.* Initial Geometry: *(b) recovered shape and appearance from a distorted visual hull (a) as initial geometry robustly recovers the shape and reflectance. In contrast, naively applying appearance-from-motion to the initial geometry fails to faithfully reproduce the appearance (c).* Capture Lighting: *(d) recovered shape and appearance from a sequence lit by a "Parking Lot" lighting environment exhibits similar quality compared to the reconstruction from a sequence captured under the "Uffizi Gallery" light probe (Figure 8.a). We assume that the incident lighting exhibits strong discontinuities; violation of this assumption results in less sharp surface reflectance estimates and a reduction in surface detail (but an overall correct global shape), as demonstrated in (e) for recovery of shape and appearance under a blurred "Uffizi Gallery" light probe (with Gaussian kernel with* $\sigma = \frac{\pi}{60} rad$*).*

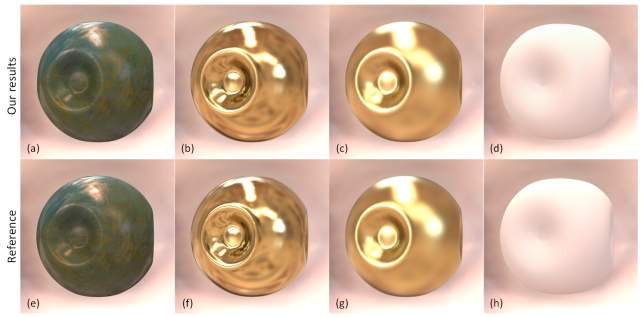

Figure 10: ***Impact of Material Properties***. *Visualizations lit by the St. Peter's Basilica light probe of recovered shape and surface reflectance with different material properties: (a) measured copper with rich texture detail, (b) glossy (roughness = 0.04), (c) rough glossy (roughness = 0.115), and (d) diffuse.*

motion [Dong et al. 2014] on the initial geometry without shape refinement. As expected, the recovered surface reflectance fails to faithfully reproduce the appearance. Similarly, directly applying a surface reconstruction method such as multi-view stereo, fails to produce reasonable geometry due to the strong specular reflections.

Impact of Material Properties We validate the impact of the underlying materials on the accuracy of the shape and appearance recovery on a selection of different material properties: a measured copper dataset [Wang et al. 2008] with rich texture detail (Figure 10.a), and three homogeneous materials (Figure 10.b-d) ranging from glossy (roughness = 0.04), rough glossy (roughness = 0.115), to diffuse – note that we do not enforce homogeneity of the appearance which is recovered for each surface point separately. Our method is able to achieve reconstructions of comparable quality for all four materials (Table 1, rows 4-7).

10.2 Robustness to Input

Impact of the Initial Geometry The initial geometry serves to indicate the size of the bounding volume, and to aid in recovering an initial lighting estimate. Therefore, our method is robust to the quality of this initial geometry, as demonstrated in Figure 9 where we recover shape and reflectance (b) from an low-frequency distorted visual hull (a). The obtained shape and reflectance is qualitatively and quantitatively comparable to the recovered results from a "clean" visual hull (compare to Figure 8.a, and Table 1, first row versus row 8).

In general, the initial shape should have the same topology as the target object, preferably smooth, as this tends to result in a smoother initial estimate of the lighting. We found that our method works well for most reasonable initial shapes. For example, all synthetic examples (excluding those in Figure 9), use a sphere as an initial shape, including the cube-example in Figure 8. We further demonstrate the robustness with respect to the initial shape on the *Clay Teacup* example, and use a sphere and the visual hull respectively as the initial geometry (Figure 11). While the quality of the recovered shape and appearance is similar, using the sphere as an initial shape requires more iterations to converge due to the larger difference in shape.

Impact of Incident Lighting We validate the repeatability of our algorithm under different types of incident lighting for the same object and material properties. Figure 9.d shows the resulting reconstruction of an object recorded under a "Parking Lot" lighting

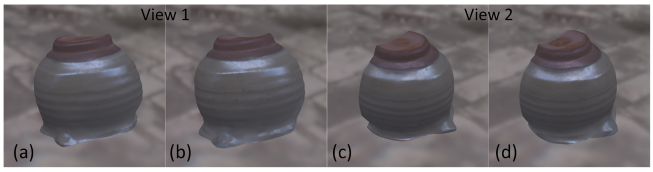

Figure 11: ***Robustness to Initial Shape***. *Visualizations from two different viewpoints of the recovering shape and appearance for two different initial shapes. Both the visual hull (a,c) and a sphere (b,d) as the initial shape produce a similar quality result.*

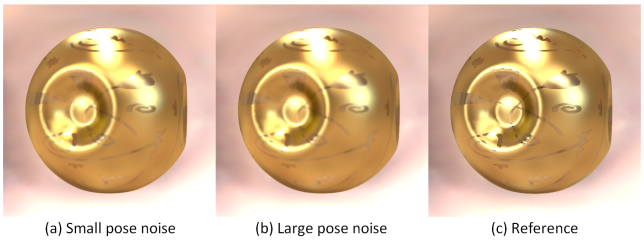

Figure 12: ***Robustness to Pose Calibration.*** *Recovered shape and appearance from image sequences where random offsets are added to the camera position and orientation. (a) reconstruction from a sequence with an average reprojection error of 0.4 pixels (2.0 maximum). (b) reconstruction from a sequence with an average reprojection error of 2.0 pixels (10.0 pixel maximum error).*

environment. The reconstruction is similar in quality to the reconstruction under the "Uffizi Gallery" light probe (compare to Figure 8.a, and Table 1, first row versus row 9). Note, the "Parking Lot" light probe exhibits a very long edge in the lighting, and it requires only three rotations to adequately "scan" each surface point, whereas the "Uffizi Gallery" light probe has relative shorter edges, and it requires at least six rotations to obtain good reconstructions. In general, smaller discontinuities require more rotations to adequately "scan" all surface points.

We rely on the presence of a sharp discontinuities in the incident lighting to infer both shape and appearance. As analyzed by Dong et al. [2014], surface reflectance will be blurred if the incident lighting does not contain such a sharp discontinuity (or if a surface point is not scanned by such an edge). Due to the close relation between surface normals, surface reflectance and observed radiance, blurred surface reflectance translates in less detailed surface normals. However, due to the general smoothness of the apparent normal field, global shape is less affected by a lack of sharpness in the incident lighting (Figure 9.e).

Impact of Object Pose Calibration Our algorithm expects for each input frame corresponding relative object pose transformations. To better understand the impact of inaccuracies in the calibration, we perform a synthetic experiment where we add random offsets to the relative camera positions and orientations. Figure 12 shows two such cases where the average reprojection error is 0.4 pixels and 2.0 pixels, and with a maximum reprojection error of 2.0 pixels and 10.0 pixels respectively. Despite the significant error in the calibration, our method is still able to produce good quality geometry and surface reflectance (Figure 12). In general, the reconstructions are more blurred than the ground truth, with a reconstruction error less than an order of magnitude larger than a well-calibrated reconstruction (Table 1, rows 10-11).

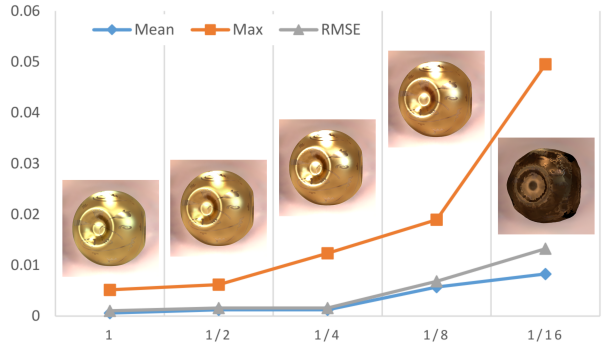

Figure 13: ***Impact of Rotation Speed.*** *Relative Hausdorff distance (normalized by the length of the largest bounding box axis) for mean, maximum and RMS error on the recovered geometry for a temporally subsampled sequence mimicking different object rotation speeds with a fixed camera exposure.*

Impact of Rotation Speed A final variable is the rotation speed at which we record the object. Assuming a fixed exposure, faster rotation implies a subsampling of the temporal traces. Hence, its effect will depend on the sharpness of the specular reflections – a low sampling rate could result in missing the passage of a lighting edge. Figure 13 shows the effect of subsampling the frame rate on the recovery of shape and appearance of the synthetic scene captured under the "Uffizi Gallery" light probe. A full sampling rate corresponds to 4320 frames or 6 full rotations at 0.5 degrees per frame. The visual quality of the recovered shape and reflectance remains good upto a $1/8$-th sampling rate, after which it degrades substantially. Quantitatively, the mean and RMS error vary little for a $1/4$-th reduction or less.

10.3 Limitations

Recovering both shape and surface reflectance under uncontrolled lighting is a challenging problem. While our method works well in many practical situations, it is not without limitations.

The first stage of our shape reconstruction algorithm works on a discrete voxel grid with finite resolution, which places an lower limit on the thickness of geometrical features that can be successfully recovered. Foregoing the voxel-based optimization, however, would necessitate repeated resampling of the mesh due to the large non-uniform deformations in the first few iterations, resulting in a detail loss comparable to the voxel grid resolution. Furthermore, moving a mesh vertex changes the apparent temporal trace, and thus apparent normal. In the first iterations (which exhibit large deformations), it is more efficient to precompute the apparent normals for each voxel rather than recompute them at every vertex deformation.

Furthermore, as our method relies on appearance-from-motion [Dong et al. 2014] to recover surface reflectance and lighting, we also share its limitations with respect to interreflections and its limitation to isotropic surface reflectance. While we take occlusions in account during shape reconstruction, we ignore them for reflectance recovery [Dong et al. 2014]. This can aversely affect the normal estimation and lighting recovery for objects dominated by diffuse or rough glossy reflections especially for concave object shapes. Furthermore, similar as in appearance-from-motion, recovering the lighting is a by-product of the recovery process, and there is no guarantee on the accuracy of the recovered lighting. Due to the bandpass behavior of the surface reflectance, the recovered lighting is generally more blurred than the ground truth lighting. As in appearance-from-motion, we shock-filter the recovered lighting, and focus on the strong discontinuities to mitigate the effects of blurring. However, for very sharp specular materials, appearance-from-motion requires very accurate geometry and pose estimates to recover the incident lighting with a sufficient degree of accuracy in order to faithfully recover the surface reflectance. Small errors in the shape or pose will induce a blurring in the recover lighting, which subsequently results in a blurred surface reflectance estimate. In the case of very specular materials, our method is still able to recover a high quality shape. However, depending on the complexity of the shape, the accuracy of the recovered geometry might not suffice to recover sharp lighting details, and thus the sharp specular surface reflectance. For example, the *"Bronze Cup"* exhibits sharp specular reflections which can still be recovered because its shape is relatively simple. However, for the spherical shape used in the synthetic validations, our method is not able to recover sharp specular surface reflectance due to the more complex geometry.

Our method assumes the lighting remains static during acquisition, which might be difficult in dynamic scenes (e.g., clouds on a windy day). Furthermore, we also assume that each surface point is "scanned" from multiple directions by a strong discontinuity in the lighting. While any type of discontinuity suffices, long edges are more efficient than point discontinuities which require more rotations to ensure sufficient coverage. Furthermore, different surface points "scan" a different subset of the incident lighting, possibly resulting in a slightly different reconstruction of the appearance. However, because neighboring surface points' temporal traces are strongly correlated, these differences in appearance reconstructions vary slowly over the surface. This effect can be observed on the *Cloissoné Bell* around the lower edge (Figure 7).

As with many methods that recover various combinations of shape, reflectance, and lighting, our formulation does not formally guarantee convergence. In practice, we found that our method typically converges to a solution that qualitatively matches the expected solution. It is an interesting avenue for future research to derive, starting from the proposed method, a formal theory of joint recovery of shape and reflectance under uncontrolled natural lighting.

Finally, our method requires prior knowledge on the object pose with respect to the camera and lighting, which might be difficult to obtain for certain cases. While our method is not married to a particular pose estimation technique, it does require high quality pose estimates, especially for objects containing sharp specular material properties.

11 Conclusion

We presented a robust integrated method for estimating both shape and appearance that only requires a video of a rotating object under unknown and uncontrolled lighting. Our method does not rely on active illumination or on specialized hardware, making it an accessible and practical method applicable under wide variety of conditions. The key enabling observation is that reflectance, surface normal, and surface position affect the temporal trace differently. We demonstrated the robustness and effectiveness of our method on a variety of synthetic and real test cases, exhibiting different combinations of shapes and material properties.

Avenues for future work include joint camera and object orientation calibration, and handling interreflections robustly. Furthermore, we would like to extend our method to the case of moving the camera instead of rotating the object.

Acknowledgements

We wish to thank Guojun Chen for rendering and video compositing, and the anonymous reviewers for their constructive feedback. Pieter Peers was partially funded by NSF grants: IIS-1217765, IIS-1350323, and a gift from Google.

References

ACKERMANN, J., RITZ, M., STORK, A., AND GOESELE, M. 2012. Removing the example from example-based photometric stereo. In *ECCV*, 197–210.

ADATO, Y., VASILYEV, Y., ZICKLER, T., AND BEN-SHAHAR, O. 2010. Shape from specular flow. *IEEE PAMI 32*, 11 (Nov), 2054–2070.

ASHIKHMIN, M., PREMOZE, S., AND SHIRLEY, P. 2000. A microfacet-based BRDF generator. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 65–74.

BARRON, J. T., AND MALIK, J. 2015. Shape, illumination, and reflectance from shading. *IEEE PAMI*.

BASRI, R., JACOBS, D. W., AND KEMELMACHER, I. 2007. Photometric stereo with general, unknown lighting. *IJCV 72*, 3, 239–257.

CHANDRAKER, M. 2014. On shape and material recovery from motion. In *ECCV*, 202–217.

COOK, R. L., AND TORRANCE, K. E. 1982. A reflectance model for computer graphics. *ACM Trans. Graph. 1*, 1, 7–24.

DONG, Y., WANG, J., TONG, X., SNYDER, J., LAN, Y., BEN-EZRA, M., AND GUO, B. 2010. Manifold bootstrapping for SVBRDF capture. *ACM Trans. Graph. 29*, 4, 98:1–98:10.

DONG, Y., CHEN, G., PEERS, P., ZHANG, J., AND TONG, X. 2014. Appearance-from-motion: Recovering spatially varying surface reflectance under unknown lighting. *ACM Trans. Graph. 33*, 6, 193:1–193:12.

FIALA, M. 2005. Artag, a fiducial marker system using digital techniques. In *CVPR*, vol. 2, 590–596.

HERTZMANN, A., AND SEITZ, S. M. 2003. Shape and materials by example: A photometric stereo approach. In *CVPR*, 533–540.

KAZHDAN, M., AND HOPPE, H. 2013. Screened poisson surface reconstruction. *ACM Trans. Graph. 32*, 3, 29:1–29:13.

LOMBARDI, S., AND NISHINO, K. 2016. Reflectance and illumination recovery in the wild. *IEEE PAMI 38*, 1, 129–141.

LU, F., MATSUSHITA, Y., SATO, I., OKABE, T., AND SATO, Y. 2013. Uncalibrated photometric stereo for unknown isotropic reflectances. In *CVRP*, 1490–1497.

NEHAB, D., RUSINKIEWICZ, S., DAVIS, J., AND RAMAMOORTHI, R. 2005. Efficiently combining positions and normals for precise 3d geometry. *ACM Trans. Graph. 24*, 3, 536–543.

NICODEMUS, F. E., RICHMOND, J. C., HSIA, J. J., GINSBERG, I. W., AND LIMPERIS, T. 1977. Geometric considerations and nomenclature for reflectance. *Monograph 161,National Bureau of Standards (US)*.

OXHOLM, G., AND NISHINO, K. 2012. Shape and reflectance from natural illumination. In *ECCV*, 528–541.

OXHOLM, G., AND NISHINO, K. 2014. Multiview shape and reflectance from natural illumination. In *CVPR*, 2163–2170.

PALMA, G., CALLIERI, M., DELLEPIANE, M., AND SCOPIGNO, R. 2012. A statistical method for svbrdf approximation from video sequences in general lighting conditions. *Comput. Graph. Forum 31*, 4, 1491–1500.

ROMEIRO, F., AND ZICKLER, T. 2010. Blind reflectometry. In *ECCV*, 45–58.

SEITZ, S. M., CURLESS, B., DIEBEL, J., SCHARSTEIN, D., AND SZELISKI, R. 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, 519–528.

TREUILLE, A., HERTZMANN, A., AND SEITZ, S. M. 2004. Example-based stereo with general BRDFs. In *ECCV*, 457–469.

TRIGGS, B., MCLAUCHLAN, P. F., HARTLEY, R. I., AND FITZGIBBON, A. W. 1999. Bundle adjustment - a modern synthesis. In *ICCV*, 298–372.

VALGAERTS, L., WU, C., BRUHN, A., SEIDEL, H.-P., AND THEOBALT, C. 2012. Lightweight binocular facial performance capture under uncontrolled lighting. *ACM Trans. Graph. 31*, 6, 187:1–187:11.

WANG, J., ZHAO, S., TONG, X., SNYDER, J., AND GUO, B. 2008. Modeling anisotropic surface reflectance with example-based microfacet synthesis. *ACM Trans. Graph. 27*, 3, 41:1–41:9.

WANG, T.-C., CHANDRAKER, M., EFROS, A., AND RAMAMOORTHI, R. 2016. Svbrdf-invariant shape and reflectance estimation from light-field cameras. In *CVPR*.

WEINMANN, M., AND KLEIN, R. 2015. Advances in geometry and reflectance acquisition. In *ACM SIGGRAPH Asia, Course Notes*.

WOODHAM, R. J. 1980. Photometric method for determining surface orientation from multiple images. *Optical Engineering 19*, 1, 3050–3068.

WU, C., WILBURN, B., MATSUSHITA, Y., AND THEOBALT, C. 2011. High-quality shape from multi-view stereo and shading under general illumination. In *CVPR*, 969–976.

WU, H., WANG, Z., AND ZHOU, K. 2016. Simultaneous localization and appearance estimation with a consumer RGB-D camera. *IEEE Trans. Vis. and Comp. Graph. 2*, 8, 2012–2023.

XU, D., DUAN, Q., ZHENG, J., ZHANG, J., CAI, J., AND CHAM, T.-J. 2014. Recovering surface details under general unknown illumination using shading and coarse multi-view stereo.

ZHANG, Z. 2000. A flexible new technique for camera calibration. In *IEEE PAMI*, vol. 22, 1330– 1334.