EUROGRAPHICS 2013 / I. Navazo, P. Poulin
(Guest Editors)

Volume 32 (2013), Number 2

Capturing Relightable Human Performances under General Uncontrolled Illumination

Guannan Li[1,2], Chenglei Wu[3,4], Carsten Stoll[3], Yebin Liu[1], Kiran Varanasi[3], Qionghai Dai[1], Christian Theobalt[3]

[1]Department of Automation - Tsinghua University, [2]Graduate School at Shenzhen - Tsinghua University,
[3]MPI Informatik, [4]Intel Visual Computing Institute

Figure 1: *Several images of a reconstructed real-world performance rendered from novel viewpoints and under a novel lighting condition. (Environment map courtesy of Paul Debevec)*

Abstract

We present a novel approach to create relightable free-viewpoint human performances from multi-view video recorded under general uncontrolled and uncalibated illumination. We first capture a multi-view sequence of an actor wearing arbitrary apparel and reconstruct a spatio-temporal coherent coarse 3D model of the performance using a marker-less tracking approach. Using these coarse reconstructions, we estimate the low-frequency component of the illumination in a spherical harmonics (SH) basis as well as the diffuse reflectance, and then utilize them to estimate the dynamic geometry detail of human actors based on shading cues. Given the high-quality time-varying geometry, the estimated illumination is extended to the all-frequency domain by re-estimating it in the wavelet basis. Finally, the high-quality all-frequency illumination is utilized to reconstruct the spatially-varying BRDF of the surface. The recovered time-varying surface geometry and spatially-varying non-Lambertian reflectance allow us to generate high-quality model-based free view-point videos of the actor under novel illumination conditions. Our method enables plausible reconstruction of relightable dynamic scene models without a complex controlled lighting apparatus, and opens up a path towards relightable performance capture in less constrained environments and using less complex acquisition setups.

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Color, shading, shadowing, and texture

1 Introduction

Capturing real performances of human actors and reproducing them in virtual environments has been one of the grand challenges in the fields of Computer Graphics and Computer Vision in the last few decades. Recent advances in marker-less multi-view video based capture methods have made it possible to reconstruct motion, geometry and texture of actors [VBMP08, GSA*09, dAST*08], and create new motions of the performances [SGdA*10] from arbitrary viewpoints. However, reconstructing a realistic *appearance* of the models is still challenging.

So far, most methods for rendering captured performances resort to projective texturing from the input video frames, e.g. [SH07, MBR*00]. Rendering a captured scene under new illumination, however, has not yet been feasible. To overcome

this limitation, some dynamic scene reconstruction methods estimate a spatially-varying BRDF of the scene model by filming such scenes under calibrated studio lighting [TAL*07]. Other approaches combine scene reconstruction with image-based relighting techniques [MPN*02, ECJ*06] by recording under an advanced controllable light stage. Despite of the great advancements in dynamic scene capture technology, these approaches are fundamentally limited by the fact that they require complex, expensive and controlled camera and light setups which only exist in controlled studio environments. For many practical animation productions in movies or games, or for recording of 3D video, the requirement of controlled calibrated lighting is highly obstructive. Essentially, movie professionals would like to capture performances that can be relit directly on an arbitrary movie set where lighting can be arbitrary and greatly vary over time. The importance of this becomes clearer if one considers that many production sets are actually outdoors.

In this paper we therefore propose a performance capture method that reconstructs detailed spatio-temporally coherent dynamic scene geometry and a spatially-varying parametric surface reflectance model of a human from a sparse set of multi-view video recordings under general uncontrolled illumination. Estimating a relightable performance under general lighting is a hard chicken-and-egg problem. In the beginning, neither shape, illumination, nor surface reflectance are known. Solving for all these high-dimensional unknowns in one go is infeasible and highly ill-conditioned, in particular if high-frequency components of illumination, geometry and reflectance are to be recovered [RH01b]. To reach a plausible solution, we therefore resort to a cleverly designed coarse-to-fine reconstruction scheme that eventually outputs highly detailed dynamic scene geometry, an all-frequency model of incident illumination, and a parametric spatially-varying BRDF model for the moving surface. The interplay of coarse-to-fine steps is designed to keep the individual sub-estimation problems feasible in terms of computation time and the signal processing theory of inverse rendering [RH01b].

Plausibly relit performances can be created from multi-view video footage under general unknown lighting. Besides enabling performance relighting in general environments, like movie sets with permanently changing surroundings, this also enables us to work with multi-view data that was captured in other labs, and for which the lighting situation was not measured.

2 Related Work

Performance Capture and Performance Editing Our method builds up on recent progress in multi-view performance capture. Several methods from that category employ vision-based 3D reconstruction, such as shape-from-silhouette or active or passive stereo [MBR*00, SH07, WWC*05]. Model-based approaches deform a shape template such that it resembles a person [dAST*08, VBMP08, GSA*09] or a person's apparel alone [BPS*08] in multi-view video, which yields spatio-temporally coherent reconstructions. The approach by [CBI10] makes a weaker *a priori* assumption by modeling the scene as a set of moving patches that are tracked over time. None of the above methods can display a human performance under new lighting conditions.

Surface and Illumination Estimation By making some prior assumptions about illumination and reflectance in a scene, the quality of 3D models reconstructed by image-based methods can be greatly improved. If illumination is controlled, higher shape detail can be reconstructed. Hernandez et al. [HVB*07] employ temporally alternating color lights for estimating detailed geometry in dynamic scenes. Vlasic et al. [VPB*09] use multi-view high-speed video captured in a controlled light-stage to reconstruct detailed geometry of a moving human by means of photometric stereo.

Even if illumination is *a priori* unknown and uncalibrated, but if certain general assumptions are met, improved 3D reconstruction is feasible. Basri et al. [BJK06] describe a single-view photometric stereo method for static Lambertian scenes using images taken under multiple lighting situations. [WWMT11] take multi-view images of a static object with constant albedo, estimate the incident illumination in spherical harmonic basis, and perform shading-based refinement based on an initial stereo reconstruction. [WVL*11] extend this approach to a spatio-temporal framework that handles moving objects with piecewise constant albedos, and [WVT12] extend it to a scenario with varying illumination. However, these approaches are limited to reconstructing approximately Lambertian surfaces.

More general reflectance models have also been considered in the context of shape refinement. Using multi-view images under known illumination, [YPS10] estimate a parametric BRDF model and exploit this for surface refinement. [Geo03] use photometric stereo to capture static shape and BRDF of a face from multiple images illuminated with a point light from unknown directions. [GCHS05] reconstruct shape and spatially-varying BRDF via photometric stereo from images under controlled lighting. Carceroni et al. [CK02] capture a moving surfel model and per-surface reflectance data of a face from multi-view video footage.

In our work we build up and extend on recent progress in shape reconstruction under general unknown illumination, in order to capture high-frequency illumination and reflectance in dynamic scenes under general lighting.

Reflectance Estimation and Relighting In the past, a variety of approaches have been proposed for image-based estimation of reflectance models for static scenes. Having a shape model, samples of surface reflectance can be recorded by capturing images of the object from varying outgoing and incident light directions with a calibrated point light. An analytical model of surface reflectance, such as a parametric BRDF, can now be estimated for the whole surface or for every surface point individually, e.g. [SWI97, LKG*03, MPBM03]. Given a shape model and some general prior assumptions about lighting [YM98], or given geometry and calibrated lighting [YDMH99], the spatially-varying BRDF of a scene can be found via inverse global illumination. Given a manually designed model of the geometry and lighting, BRDF estimation from a single image is feasible [BG01]. [TAL*07] extend this concept to scenes with a moving human. They reconstruct a shape and motion of the actor using template-based motion estimation approach from multi-view video recorded under the light

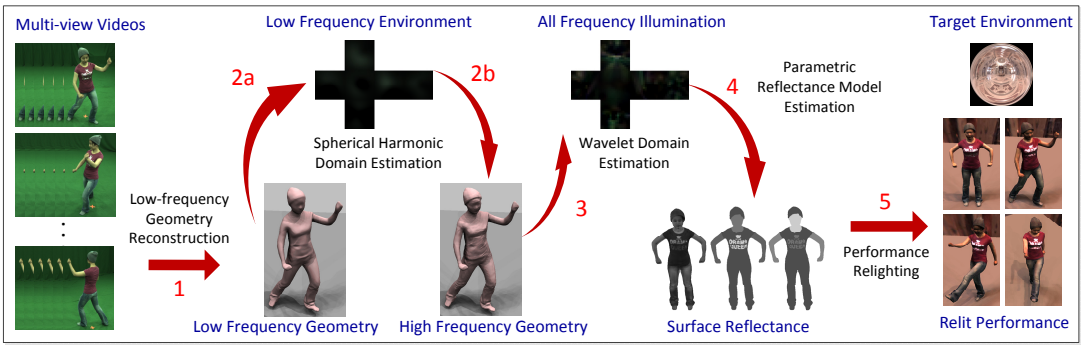

Figure 2: *Overview of the method, illustrating the steps for geometry reconstruction (1 and 2b), lighting estimation (2a and 3), reflectance estimation (4), and final performance relighting (5).*

of two calibrated spot lights. From the data, they estimate a parametric BRDF model for each vertex. However, their 3D models are very coarse which has a negative influence on the final result.

An alternative approach to relighting is image-based methods that do not reconstruct an analytical reflectance model but perform relighting by proper combination of a set of input images taken under different illuminations. [DHT*00] capture facial reflectance fields using a controlled light stage. Einarsson [ECJ*06, WGT*05] extend this concept to dynamic scenes, using a fast light stage and high speed cameras for recording. However, the complexity of their setup restricts relighting to a single camera viewpoint [WGT*05], or requires the motion to be simple and periodic if a viewpoint change is desired [ECJ*06]. In contrast, our algorithm can estimate relightable versions of general human performances under general uncontrolled lighting that can be displayed from arbitrary viewpoints.

An important component of our approach is an algorithm for estimating the incident illumination based on a refined estimation of scene geometry. For this purpose, we need to properly parameterize the environment map of incident illumination. Low-frequency illumination can be efficiently represented using Spherical Harmonics (SH), as was shown by Ramamoorthi and Hanrahan [RH01a] and several other papers. The SH basis has also been used in the signal processing theory of inverse rendering [RH01b]. Spherical harmonics can only represent low frequency illumination or reflectance effects; aliasing occurs around high-frequency illumination or reflectance effects. Hence, Ng et al. [NRH04] proposed to use a Haar wavelet basis to model high-frequency lighting and reflectance effects. Using wavelet-based lighting and an assumed subspace of BRDFs, the surface reflectance of a static object can be estimated using images from community image databases [HFB*09]. We also subsequently solve a variety of inverse rendering problems and combine the advantages of the spherical harmonics representation and the wavelet representation for incident illumination in a coarse-to-fine strategy.

3 Overview

Input to our system is a multi-view video sequence of a moving actor captured using a sparse set of N_c synchronized cameras running at standard frame rate (N_c typically between 8 and 9). The cameras are expected to be geometrically calibrated. They are also assumed to have linear response (if exact response curves are available, they are used), and color matching across views is done during pre-processing. However, we do not impose strict requirements concerning the scene illumination.

Given this input data our algorithm reconstructs high resolution spatio-temporally coherent geometry, surface reflectance and incident illumination. As estimating all parameters simultaneously in high accuracy would be too difficult, we gradually refine the estimations over several steps of the pipeline as shown in Fig. 2. Firstly, we reconstruct a faithful yet relatively coarse spatio-temporally coherent model of an actor from multi-view video (Fig. 2 step 1, Sec.4.1). Secondly, using this coarse scene geometry, we reconstruct a coarse estimate of *low-frequency* diffuse surface reflectance and incident lighting in spherical harmonics. Given an estimate of diffuse material and low-frequency lighting, we can spatio-temporally refine the surface geometry to recover previously missing fine-scale shape detail (Fig. 2 step 2, Sec.4.2). Thirdly, we use the now available more detailed high-frequency shape model to compute an all-frequency representation of the incident illumination in the wavelet domain (Fig. 2 step 3, Sec.5). Finally, we use the high-frequency illumination model to estimate a spatially-varying parametric BRDF model for the dynamic shape which can also represent high-frequency reflectance effects (Fig. 2 step 4, Sec.6).

4 Geometry Reconstruction

To keep the surface reconstruction problem tractable, we first track a smooth template to obtain the coarse *low-frequency* mesh reconstructions for each time frame, Sect. 4.1. Then we perform spatio-temporal surface refinement at each frame to obtain the *high-frequency* temporally varying geometry component, which is similar to [WVL*11] but develops a new spatio-temporally coherent material segmentation method.

4.1 Low-frequency Geometry Reconstruction

We reconstruct an initial low-frequency estimate of the time-varying geometry of the actor using the performance capture method of [GSA*09]. It uses an initial shape template of the human that comprises of a surface mesh M and a kinematic skeleton with associated skinning weights. In our case, the

surface mesh is obtained from a static laser scan, but can also be obtained via image-based reconstruction from the multi-view image data. Our surface meshes typically have around $N_v = 80000$ vertices. We purposefully smooth out high-frequency surface geometric detail from the mesh before tracking to ensure that we only reconstruct low-frequency components at this stage. The performance capture algorithm automatically reconstructs the initial set of coarse surface meshes M_c^t for each time-step t. They represent the coarse motion and shape of the actor, but lack high-frequency surface detail, Fig. 3(b).

4.2 High-frequency Geometry Reconstruction

To recover fine-scale time-varying scene geometry, we employ a variant of [WVL*11]. We use the coarse scene geometry M_c^t and the input video frames to estimate the incident illumination and a coarse piecewise-uniform surface albedo map at each time-step. In this step, we assume that the surface reflectance at each point on the surface is Lambertian (a constraint we will relax at a later stage in our pipeline) and that incident illumination can be represented using a low-order SH representation. Also, even though the diffuse surface albedo may vary arbitrarily in general scenes, in most cases it is fair to assume that there is a finite number of basis materials [HFB*09, TAL*07].

We assume the surface albedos belong to a set of k distinct materials $\{d_1^t, \ldots, d_k^t\}$ and want to determine the diffuse albedo labels $\boldsymbol{a}^t \in \{1, \ldots, k\}$ of the vertices. In [WVL*11], the material segments are obtained by a graph-based segmentation method before estimating lighting and albedo [FH04]. While the diffuse segmentation obtained with their approach suffices for the purpose of shape refinement, we require more accurate albedo estimates that are also suitable for our parametric reflectance estimation, where preservation of spatio-temporally coherent material boundaries is critical. We therefore develop a new method that generates a segmentation which uses a consistent set of materials for each time-step, preserves boundaries between materials on the surface, and is able to represent the potential shifting of material over the surface. This is caused for example by shifting apparel (Fig. 3(c)). Assuming the material segmentation $\boldsymbol{a}^{t-1}$ of the previous frame is given, we segment the mesh of the current frame by finding the least energy configuration of the Markov-Random-Field (MRF) defined as:

$$\psi(\boldsymbol{a}^t) = \sum_{i \in M_c^t} (\phi(O|a_i) + \sum_{j \in N(i)} \phi(a_i, a_j)), \qquad (1)$$

where $N(i)$ is the neighboring vertex set of vertex i, $\phi(a_i, a_j)$ is a smoothness term that takes the form of a generalized Potts model [SZS*08], and $\phi(O|a_i)$ is a likelihood data term which impose individual penalties for assigning a albedo label to vertex i according to the observation O. The data term combines two terms: The first one penalizes different albedo color from the assigned material, and the second term is the albedo label prior, which penalizes different labels in consecutive time-steps. The MAP-MRF energy function in Eq. 1 is minimized via graph cuts [SZS*08]. Given the labels $\boldsymbol{a}^t \in \{1, \ldots, k\}$ of the mesh at time t, we now solve a global MAP inference problem that updates the albedo values $\{d_1^t, \ldots, d_k^t\}$ for each label and

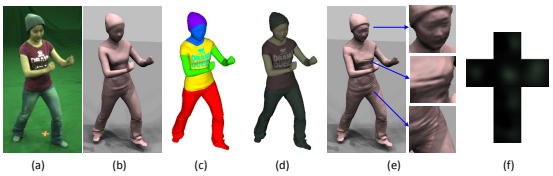

Figure 3: *Geometry Reconstruction. (a) Captured image, (b) coarse geometry from performance capture, (c) material segmentation, (d) average albedo map, (e) refined geometry, (f) computed environment map using SH.*

simultaneously estimates the incident illumination using a SH basis (Fig. 3 (d) and (f)), as described in [WVL*11].

The albedo estimation is performed for each RGB color channel. Considering that the color ambiguity between lighting and albedo can not be totally solved without other assumptions, we initially assume monochromatic illumination and solve for monochromatic incident illumination. As this may have an influence on later stages in the pipeline, we ask the user to provide a reference color for single visible material patch (usually gray), and update the SH light color accordingly (similar to manual white balancing).

Having estimated illumination and albedo, the coarse geometry of each frame is now refined by solving another spatio-temporal MAP inference problem as described in [WVL*11]. The result of fine geometry estimation are not only the per time-step refined meshes M_r^t with refined geometry $\boldsymbol{g}^t$, but also a set of piece-wise uniform diffuse albedo maps $\boldsymbol{d}^t$ and incident illumination estimates $\boldsymbol{l}^t$ that we use to initialize the all-frequency lighting estimation (Sec. 5) and BRDF estimation (Sec. 6).

5 All-frequency Lighting Estimation

Given a model of shape, illumination and surface reflectance, the rendering equation describes the local surface light transport [Kaj86]. If incident illumination is given in discretized form as an environment map l with i entries, this equation takes the form:

$$I(x, \omega_o) = \sum_i f_r(x, \omega_o, \omega_i) v(x, \omega_i) l(\omega_i)(n_x \cdot \omega_i), \qquad (2)$$

where ω_i and ω_o are the incoming and outgoing light directions at surface location x. $I(x, \omega_o)$ is the outgoing radiance in direction ω_o. f_r is the bidirectional reflectance distribution function (BRDF) for the incident illumination from direction ω_i at position x and the outgoing direction ω_o [HFB*09], v is the visibility map of point x, $l(\omega_i)$ is the light intensity of the environment, and n_x is the vertex normal.

In order to estimate non-Lambertian high-frequency reflectance properties of the surface, it is necessary to know the full-frequency incident illumination. However, the SH illumination estimated in Section 4.2 only represents a low-frequency approximation. To overcome this limitation, we refine the illumination estimation in each color channel using a wavelet representation that can model all-frequency lighting.

We combine the BRDF $f_r(x, \omega_o, \omega_i)$, visibility functions

$v(x,\omega_i)$, and the cosine term $(n_x \cdot \omega_i)$ from Eq. 2 into $T(\omega_o)$, and rewrite the light transport in the matrix form:

$$I(\omega_o) = T(\omega_o)L, \tag{3}$$

where $T(\omega_o)$ is the two dimensional $N_v \times m$ transport matrix of the captured surface under view direction ω_o, and $I(\omega_o)$ and L are column vectors of sizes N_v and m for radiance intensity and incident illumination. Here, m is the number of pixels in the environment map. Estimating the incident illumination directly is untractable due to the large number of coefficients to be estimated. Instead, we now use a 2D Haar wavelet basis W [NRH04] to represent the transport matrix and simplify the problem:

$$I(\omega_o) = T_w(\omega_o)L_w, \tag{4}$$

where

$$T(\omega_o) = T_w(\omega_o)\cdot W^T, L = W \cdot L_w. \tag{5}$$

In the Haar wavelet domain even high-frequency illumination effects can be represented using a much less number of coefficients, thus $dim(L_w) << dim(L)$. This keeps the inverse rendering problem of solving for the incident lighting tractable. Using the new basis, we estimate the all-frequency lighting environment as the minimizer L_w of the energy:

$$E_L = \sum_{j \in N_v} \sum_{c \in \mathcal{V}(j)} \Big(\big\| I(F_c(i)) - T_{w,j}(\omega_{o,j}(c))L_w \big\| + \gamma_1 \| W \cdot L_w - S \cdot L_{SH} \| + \gamma_2 \| TV(W \cdot L_w) \|_1 \Big). \tag{6}$$

Here, $I(F_c(j))$ denotes the captured radiance of vertex j as seen from camera image F_c, and $\omega_{o,j}(c)$ is the outgoing light direction from the vertex to the pixel $F_c(j)$ in camera c. $T_{w,j}(\omega_{o,j}(c))$ is the transport matrix for vertex j and outgoing direction $\omega_{o,j}(c)$. $\mathcal{V}(j)$ are the cameras in which surface point j is visible. S is the SH basis matrix defined in the spatial domain. L_{SH} is the spherical harmonic coefficients of the illumination estimated previously, and $TV(\cdot)$ is the total variation of the spatial environment map.

Estimating the illumination from the images using only the first term in Eq. 6 is insufficient, as in the general case the problem is ill-posed, and noise may have a starkly deteriorating influence [RH01b]. Because of this we include the previously estimated SH basis illumination L_{SH} as regularizer, and add a smoothness constraint on the environment map. To prevent oversmoothing of the lighting estimation, we minimize the total variation [Li11], which preserves the high-frequency structure of the lighting but prevents noise. For incident illumination, we solve the intensity in each color channel respectively. Please see the supplementary documents for extensive comparisons with other lighting estimation methods.

Minimization In practice, we do not reconstruct a wavelet lighting estimate for every time-step of video, but for a representative subset of time-steps. This is a valid compromise between computation time and potential temporal or location-dependent variation in incident illumination, as discussed in more detail in Sec. 6.

So far, we have not defined what BRDF f_r to use when minimizing Eq. 6. From low-frequency geometry and lighting estimation, we merely have a diffuse estimate of surface reflectance. Using that diffuse albedo distribution directly, however, is not ideal. It mostly captures low frequency effects and has low-bandwidth, which limits the high-frequency illumination effects that can be estimated [RH01b]. To capture high-frequency illumination, we would ideally like to assume a full Phong model with diffuse and specular (high-frequency) components, for its compact representation while maintaining a reasonable accuracy. Up-to-now this model is not available. To solve this chicken-and-egg problem, we solve for L_w in two steps:

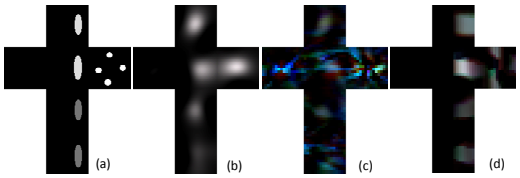

Figure 4: *Estimated lighting. (a) Ground-truth environment map, (b) low-frequency estimate in spherical harmonics, (c) all-frequency estimate assuming diffuse BRDF, (d) our approach: all-frequency estimation using full Phong BRDF.*

First, we use the approach from Sec. 6.1 to estimate a set of per-segment Phong BRDF parameters on the surface $M_r(t_r)$. This also gives us an initial coarse estimate of high-frequency specular components. The reason for this is that we consider visibility, and thus local shape variation and occlusion, during low-frequency lighting estimation. Therefore, the SH representation also encodes some higher frequency effects that are needed to capture at least a rough estimate of specular reflectance. We now initialize the solution of Eq. 6 by using those BRDFs to set up the transport matrices. This yields a faithful reconstruction of high-frequency lighting. Fig. 4 validates these steps on a synthetic sequence (see also Sec.8). While direct reconstruction of wavelet lighting with a diffuse BRDF (Fig. 4(c)) already leads to improved incident lighting estimates compared to SH estimation (Fig. 4(b)), our two-step procedure (Fig. 4(d)) with the intermediate reconstruction of a Phong model leads to even better results with better localized light source distributions. Iterating this two-step procedure may lead to further improvements. However, in our experiments we found these to be marginal. Therefore, we typically resort to only one iteration.

6 Parametric Reflectance Reconstruction

Given the estimate of the all-frequency illumination and detailed time-varying geometry, we now estimate the non-Lambertian reflectance of each surface point. To this end, we estimate for each location x on the mesh surface at each time-step the parameters of a Phong BRDF model,

$$f_r(x,\omega_o,\omega_i) = \rho_d^x + \frac{\rho_s^x}{n_x \cdot \omega_i}(r_{x,i}^x \cdot \omega_o)^{\alpha^x}, \tag{7}$$

where ρ_s^x is the specular albedo, α^x the specular exponent, and $r_{x,i} = 2(n_x \cdot \omega_i)n_x - \omega_i$ is the reflected direction of ω_i about n_x. The Phong model features a low-frequency diffuse component and a specular lobe to model high-frequency reflectance. We selected the Phong model since it enables us to represent

diffuse and specular surface reflectance based on a small set of parameters. It is thus a good compromise between modeling strength and computational complexity.

The Phong BRDF model requires us to estimate three material parameters for each point on the surface of the object, namely diffuse albedo ρ_d^x, specular albedo ρ_s^x, and a specular exponent α^x. Optimization of these parameters from a single set of multi-view images of a certain time-step is in general an under-determined problem since we would not acquire enough reflectance samples for each surface point. While estimation of the diffuse low-frequency reflectance is in most cases feasible from a single time-step of video, this does not hold for the high-frequency component where more samples are required. To solve for the reflectance at each surface point, we make use of the following assumptions:

The surface of a human actor usually does not contain an arbitrary number of materials, but rather a set of base materials. Specular albedo and shininess will be very similar in each material segment and not vary over time, while the diffuse albedo will exhibit stronger variations in a segment due to small alignment errors and micro-scale surface detail that is not captured by our surface geometry and may also change over time. This observation was also used in some other previous work, e.g., [LKG*03, YDMH99]. We use the material segmentations obtained for each time-step in Sec. 4 to identify the set of base materials, and their potential variation in spatial layout over time. By this means, we can also represent certain temporal variation, such as the shifting and stretching of clothing on the body. Further on, we have full temporal correspondence for each vertex over time, and the actor is moving in the scene relative to camera and lighting, we can expand the set of samples by accumulating samples over time for each surface point.

Based on these assumptions, we divide the reflectance reconstruction into two different steps. First, we estimate a single average BRDF for each material segment using a subset of the time-steps of the input sequence. Then, based on the average BRDF parameters of each segment, we update the diffuse albedo of each vertex for each time-step to allow more accurate reconstruction of potential dynamic effects, e.g., due to cloth shifting, facial expressions or moving micro-structure.

6.1 Per-segment Reflectance Estimation

We first estimate a complete Phong BRDF for each material segment, using $O_r = 3$ reference time-steps out of the entire length of the input video sequence. Subsampling is necessary as estimating the parameters on all input frames would be too expensive. The time-steps are selected such that the poses of the actor across them are sufficiently different, and the actor's positions relative to the cameras are sufficiently varying. This way, an expressive set of reflectance samples is assembled. The time-steps are currently selected manually, but a simple automatic procedure would also be possible. We estimate the reflectance parameters ρ_d^K, ρ_s^K and α^K for each segment $K \in \{\mathbf{K}\}$ by minimizing the following energy function:

$$E_K = \sum_{f \in O_r} \sum_{x \in K, c \in N_c} \frac{1}{N_v(K)} \| I(\omega_{o,x}(c,f), \rho_d^K, \rho_s^K, \alpha^K)$$

Figure 5: *Average BRDF per material segment. (a) Input image, (b) diffuse albedo, (c) specular albedo (both for red channel), (d) specular exponent. All values are scaled to* $[0.3, 0.8]$*, and shown in gray scale. Note the plausibly captured high specularity of the print on the t-shirt.*

$$- F_c(x,f) \| + \lambda \left\| \rho_d^K - \bar{\rho}_d^K \right\|. \quad (8)$$

Here $I(\omega_{o,x}(c,f), \rho_d^K, \rho_s^K, \alpha^K)$ is the result of evaluating the rendering equation (Eq. 2), at vertex x at time-step f with the current BRDF parameters. $\omega_{o,c}(x,f)$ is the outgoing light direction from x towards camera c at time-step f. $I(\omega_o(x,f), \rho_d^K, \rho_s^K, \alpha^K)$ is dependent on the high-frequency incident illumination estimate $L_w(f)$ that we reconstruct for each time-step f separately using the method described in Sec. 5. This way we can cope with certain temporal changes in the illumination, but we can also compensate for the fact that light sources are not infinitely far away from the actor in indoor environments; thus lighting environment maps are position-dependent. $N_v(K)$ is the number of vertices assigned to material K, and $F_c(x,f)$ is the radiance of surface point x at time f as measured from camera image F_c. Please note that we only consider camera images in which a vertex is visible, which was left out of the above equation for better readability. $\bar{\rho}_d^K$ is the average of the diffuse reflectance values which we recovered earlier for material K during low-frequency lighting estimation and geometry refinement, Section 4.2. As stated before, the estimation of reflectance parameters is influenced by the quality of the available samples and the accuracy of the estimated incident lighting. In either of them, inaccuracies and noise may exist. We therefore use $\bar{\rho}_d^K$ as a regularizer since the diffuse albedos found during low-frequency lighting estimation can serve as a guideline. λ is a weighting factor. We use the Levenberg-Marquardt algorithm to optimize Eq. 8. To ensure that the estimated albedo values are not negative, we rewrite

$$\rho_d^K = (\beta_d^K)^2, \rho_s^K = (\beta_s^K)^2, \alpha^K = (\beta^K)^2. \quad (9)$$

The end result is an average, non-time-varying BRDF parameter set for each material (Fig. 5).

6.2 Per-frame Diffuse Update

Based on the per-segment reflectance estimation, we update the diffuse albedo for each time-step to obtain a spatially and temporally varying reflectance. Given the per-material reflectance parameters, we first render the surface under the captured environment, and then minimize the difference from image-based intensity value to optimize the spatial-varying diffuse component.

To ensure that our albedo values are spatio-temporally

consistent, we include information from a window of T neighboring frames into the optimization. This formulation over the temporal domain also lets us estimate albedo values for surface points that are not visible in any of the cameras in the current time-step (see video for visualization of the dynamic diffuse albedo). The energy function to update the diffuse albedo $\rho_d^x(t=f)$ at time-step f and at surface point x is:

$$E_x = \sum_{t=f-T}^{f+T} \sum_{c \in N_c} \beta(x,t,\omega_o,f) \| I(\omega_{o,x}(c,t), \rho_d^x(f), \rho_s^x, \alpha^x) - F_c(x,t) \| + \lambda \left\| \rho_d^x(f) - \rho_d^K \right\|, \quad (10)$$

where $\beta(x,t,\omega_o,f)$ is a weighting factor to adjust the influence of frame t from view point ω_o. We use ρ_d^K - the previously estimated diffuse albedo per-material K as a regularizer. The weighting factor $\beta(x,t,\omega_o,f)$ allows us to control the influence of the input images on the estimation. It penalizes regions with high occlusions in the geometry (measured by the ambient occlusion term $\gamma(x,t)$), as these are usually dark and noisy. On the other hand, we want to give higher weights to image points directly facing the camera, as they are more robust to small errors in the reconstructed geometry. Hence, the weighting factor takes the form:

$$\beta(x,t,\omega_o,f) = (\omega_o \cdot n_x(t)) \cdot (1 - \gamma_x(t)). \quad (11)$$

7 Performance Relighting

The final result of our reconstruction pipeline is a sequence of high resolution spatio-temporally coherent triangle meshes, including an estimated BRDF for each surface point at each point in time, and the incident illumination of the recorded performance. Please note that the BRDF data are temporally-varying for two reasons: the spatially-varying distribution of materials (even though per-material specular reflectance stays the same), and a per-time-step re-estimation of the low-frequency illumination. We can now use this data to create novel virtual performances under an arbitrary new environment lighting l_{new} and (potentially moving) camera c_{new}.

To ensure interactive performance for our renderer, we use the median-cut algorithm [Deb06] to create n importance sampled directional light sources that approximate the environment map l_{new}. We render the geometry under each light source in parallel using a GPU based renderer for a given camera c_{new}, using shadow mapping to generate shadows; and combine the images into a final relit image.

8 Results and Discussion

We reconstructed geometry, reflectance and illumination from three multi-view video sequence (see Tab. 1 for sequence lengths). The sequences *kungfu* (Fig. 1 and Fig. 8, row 1) and *dance* (Fig. 8, row 2) were reconstructed with 9 cameras in our multi-view video studio (more results can be found in the supplementary document). We used general arbitrary studio lighting which was not controlled or designed in any specific way. Cameras recorded at a resolution of 1296×972 pixels, and at a frame rate of 45 fps; they were placed in a roughly circular

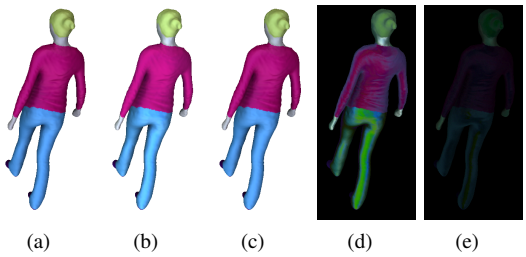

Figure 6: *Quantitative evaluation. (a) Relighting result (R_{Ref}) under a novel illumination using ground-truth BRDF parameters, (b) SH based BRDF relighting R_{SH}, (c) wavelet based BRDF relighting R_{full}, (d) difference image between R_{Ref} and R_{SH}, (e) difference image between R_{Ref} and R_{full} (difference images scaled by a factor of* 4 *for better visibility).*

arrangement around the scene. The *samba* multi-view video data set was downloaded from the web [dAST*08] (see supplementary document and video). Its recording was thus not under our control, and no information about incident illumination for that particular scene is available. The sequence is shot from 8 cameras running at 24 fps and at a resolution of 1004×1004 pixels.

The results show that our algorithm is able to capture geometry, reflectance and illumination in a believable way. Plausibly relit performances can be rendered from arbitrary new viewpoints and under novel environment lightings. Please note that for the *samba* sequence we reconstruct geometry with our own approach and do not use the geometry provided by [dAST*08] since their geometry reportedly does not reconstruct the true high-frequency shape detail. In all our results, small-scale time-varying surface detail is plausibly captured and relit, such as folds in clothing (Fig. 8 bottom row). In particular, high-frequency reflectance effects are captured, as for example the specularity in the print of the t-shirt in the *kungfu* sequence, and the slight specularities in the skin and other fabrics in the *dance* sequence (see also Fig. 5). Our approach does not assume static per-vertex reflectance, and can thus also handle changing facial expressions or shifting apparel to a certain extent. Also from the *samba* sequence which was captured outside of our lab, plausible relightable performances can be reconstructed. It is essential to watch the supplemental video to see the results in motion.

Performance Reconstructing all data of a single sequence comprising 945 frames and 9 input cameras takes 118.2 hours using a single threaded unoptimized implementation on a standard PC (timings for further sequences can be found in Table 1). Low-frequency geometry estimation (step 1) requires $\sim$ 1 minute per frame, high-frequency geometry estimation (including SH illumination estimation) $\sim$ 5 minutes per frame, wavelet lighting estimation $\sim$ 6 minutes per frame (only for representative frames), and final BRDF reconstruction $\sim$ 1.5 minutes per frame. The rendering of our reconstructed sequences using a simple GPU based renderer with a median-cut of 256 light sources runs at $\sim$ 4 frames per second on a modern desktop graphics card.

Quantitative evaluation To evaluate our approach quan-

titatively, we generated a 100 frame synthetic sequence of a 3D model from 10 virtual cameras under time varying illumination. We manually specified BRDF parameters for the surface, so that ground truth R_{ref} would be available. Using this synthetic sequence, we analyze the importance of our various algorithmic steps in the following (see also Sec. 5 and Fig. 4).

To emphasize the importance of reconstructing high-frequency incident illumination (see Sec. 5), we compare our full all-frequency reconstruction pipeline R_{full} against Phong BRDF reconstruction that uses only low-frequency SH-based illumination R_{SH} estimation. The reconstructed BRDF parameters of R_{SH} have an average error of $(0.0392, 0.0591, 23.846)$ for ρ_d, ρ_s, and α respectively, while BRDF parameters for R_{full} are estimated more accurately with an average error of $(0.0119, 0.0311, 4.147)$ (please note that the domain of ρ_d and ρ_s is $[0,1]$, while the values of α can get much larger). The biggest improvement can be found in the more accurate estimation of the specular components, which is very important for a plausible visual appearance and is one of the major benefits of our approach.

To evaluate the perceived difference, we also relight the reconstructed models under a new lighting using the estimated Phong BRDF parameters and compare them against the ground-truth. The average pixel difference in grey scale of R_{SH} and R_{ref} at 0.0623 is higher than the corresponding value for R_{full} at 0.0168, where the intensity range is $[0,1]$. This reduction in average error results in more accurately relit specular highlights on the surface (Fig. 6). For this sequence, we estimate the all-frequency illumination for each frame respectively to handle the time varying illumination. Please refer to the video and supplementary document for further results.

Ground-truth comparison on real world data We also validated the accuracy of our reconstruction and relighting approach on real world data. We captured a performer and the incident illumination in two different environments, Fig. 7(a) and (d). Using our full pipeline, we reconstructed the surface geometry, BRDF and incident illumination from the input images under the first lighting condition, Fig. 7(a), and relit the estimated geometry with the ground-truth environment map, Fig. 7(b). Up to some blurring introduced by small calibration and geometry inaccuracies, the relit result closely resembles the original image.

We then relit the reconstructed model using a different captured environment map, Fig. 7(c), for which we also captured a ground truth image of the performer in a similar pose, Fig. 7(d). Please note that fine surface details such as wrinkles will be different in images (c) and (d) as the 3D geometry was reconstructed from the input pose in Fig. 7(a), which is similar, but not identical to Fig. 7(d). Note that we purposefully did not apply our full reconstruction pipeline, to the sequence in

Sequence	Frames	Cameras	Processing Time
kungfu	945	9	118.2h
dance	460	9	57.6h
samba	192	8	25.7h

Table 1: *Details on each reconstructed sequence.*

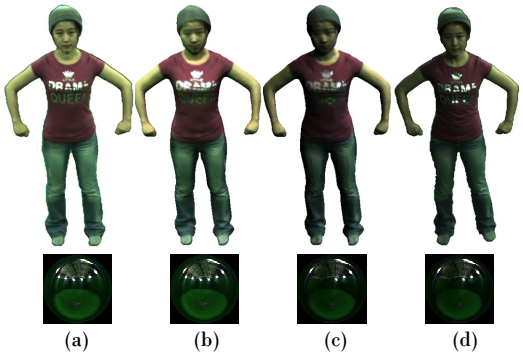

Figure 7: *Ground-truth comparison: (a) Input image captured under environment. (b) Reconstruction relit with environment. (c) The same reconstruction as (b) relit with environment. (d) Ground truth camera image caputred with environement.*

Fig. 7(d), but estimated the geometry and reflectance entirely from Fig. 7(a). We then only matched the model's pose to prevent any bias. Despite these unavoidable differences in geometry, the relit result is very similar to the ground truth image, showing that our algorithm captures the surface reflectance of a scene accurately and generalizes to other illuminations.

Further evaluation We experimentally validated all steps of our pipeline and confirm that our high-frequency reflectance and lighting estimation provides the best results (see supplementary material for details). We are able to reconstruct higher quality geometry and lighting compared to [WVL*11], who only estimate diffuse materials and low-frequency illumination (see supplementary document Sec. 2). We also validated that it is feasible to apply our pipeline to scenes with time-varying lighting (see supplementary material Sec. 1). Allowing temporally varying incident illumination also allows us to compensate for small variations in illumination in the recording volume, which may be cause due to the assumption of infinitely distant lighting.

8.1 Discussion and Future Work

Even though our full pipeline cannot guarantee fully-accurate reconstructions of geometry, reflectance, and illumination, our results nevertheless are plausible and are of a high visual quality. The geometry and reflectance are spatio-temporally coherent, and the estimated illumination captures the main components of the real ground-truth illumination.

However, the approach still is subject to some limitations. Due to calibration inaccuracies in the cameras, noise in the input images, and simplified assumptions in the reconstruction process, our geometry will not be accurate up to millimeter scale. If errors are larger, this may lead to ghosting artifacts in the reconstruction. We are also limited by the resolution of the input videos. We cannot reconstruct reflectance and geometry smaller than a pixel's size. This also limits the image resolution and zoom level at which performances can be rendered convincingly. Super-resolution approaches may be feasible and we plan to investigate this in the future.

Our reconstruction currently considers only direct illu-

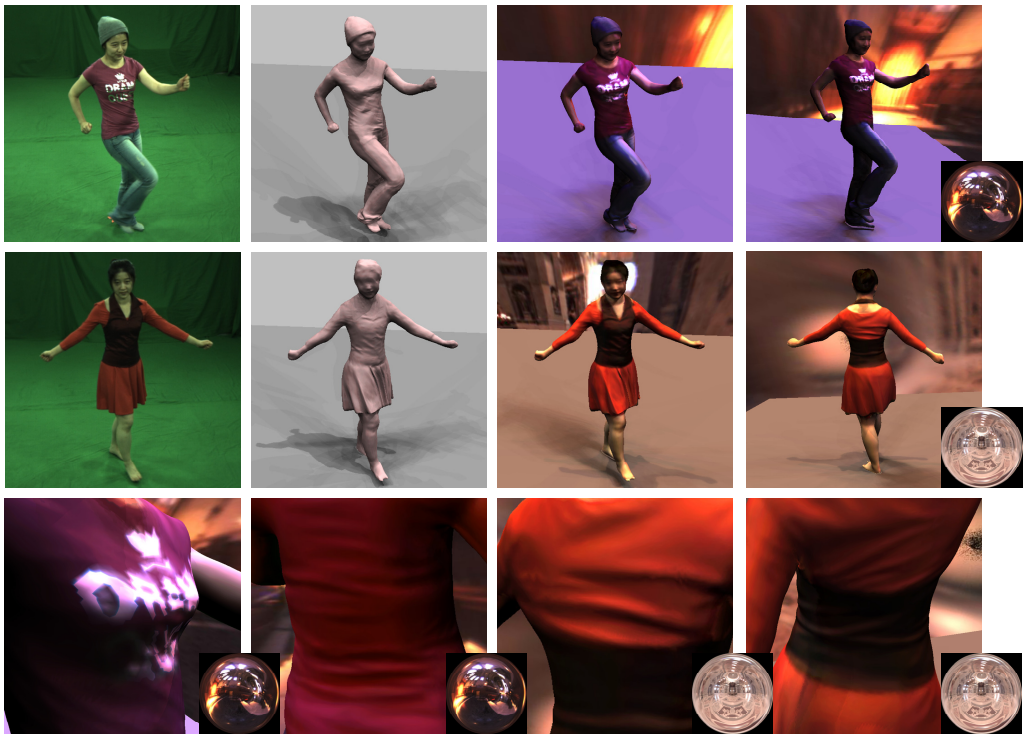

Figure 8: *Results of our algorithm from different input scenes. The first two rows show from left to right: Single input camera view, reconstructed geometry, relit performance from input camera view, and relit performance from novel view. The bottom row shows closeups on the characters, highlighting the reconstructed and relit fine geometry details and specular reflections (Environment maps Grace Cathedral and St. Peter's courtesy of Paul Debevec).*

mination coming from infinitely far light sources. This is in general not true, but the negative influence of this assumption on the reconstructions is reduced by considering local visibility and the ambient occlusion term in the BRDF estimation. We acknowledge that more advanced segmentation approaches exist that could be used to separate local and global effects. Estimating lighting for each color channel may lead to overfitting, as it is difficult to separate albedo from illumination color. However, in our experiments the system was able to separate the colors accurately on both real and artificial sequences (see supplementary material). The Phong model used in our BRDF estimation is not suitable for all input materials. For example, the appearance of human skin exhibits subsurface-scattering effects which cannot be captured by a simpler model and may lead to reconstruction errors. We plan to investigate more complex BRDF models in the future.

Our estimation is dependent on acquiring a sufficient variety of appearance samples for each surface point under varying incident lighting and outgoing viewing direction. We achieve this through the motion of the model relative to the environment. However, if the performer's motion is too limited or certain parts of the body are always in shadow, the reflectance estimation will be inaccurate. Finally, the performance capture approach we currently use [GSA*09] expects background segmentation and may not directly work in arbitrary environments. However, we believe it can be extended to outdoor environments.

9 Conclusions

We presented an approach to capture relightable human performances from sparse multi-view video footage that was recorded under uncontrolled illumination. By capturing scene geometry coarsely, and subsequently solving a sequence of carefully-designed inverse rendering problems, we are able to capture highly detailed dynamic shape, high-frequency scene illumination, and detailed spatially and temporally varying surface reflectance. Our captured performances can be plausibly rendered from arbitrary new virtual viewpoints and under arbitrary new incident lighting. In addition to the theoretical insights, our algorithm has several advantages from an application point of view: It enables relightable performance capture without complex controllable light setups, and can be applied to multi-view video captured in other labs. Even though all our examples were captured indoors, we believe that our work paves the way for relightable performance capture on outdoor sets.

10 Acknowledgements

The majority of this work was done when the first author was visiting MPI Informatik. This work has been developed within the Max-Planck-Center for Visual Computing and Communication collaboration, and has been co-financed by the Intel Visual Computing Institute. Support in China came from

the National Basic Research Project (No.2010CB731800) and the Project of NSFC (No. 60932007 & 61073072 & 61035002).

References

[BG01] BOIVIN S., GAGALOWICZ A.: Image-based rendering of diffuse, specular and glossy surfaces from a single image. In *Proc. SIGGRAPH* (2001), ACM, pp. 107–116. 2

[BJK06] BASRI R., JACOBS D., KEMELMACHER I.: Photometric stereo with general unknown lighting. *IJCV 72*, 3 (2006), 239–257. 2

[BPS*08] BRADLEY D., POPA T., SHEFFER A., HEIDRICH W., BOUBEKEUR T.: Markerless garment capture. *ACM Trans. Graphics (Proc. SIGGRAPH) 27*, 3 (2008), 99. 2

[CBI10] CAGNIART C., BOYER E., ILIC S.: Free-form mesh tracking: a patch-based approach. In *Proc. IEEE CVPR* (2010). 2

[CK02] CARCERONI R. L., KUTULAKOS K. N.: Multi-view scene capture by surfel sampling: From video streams to non-rigid 3d motion, shape and reflectance. *Int. J. Comput. Vision 49* (September 2002), 175–214. 2

[dAST*08] DE AGUIAR E., STOLL C., THEOBALT C., AHMED N., SEIDEL H.-P., THRUN S.: Performance capture from sparse multi-view video. *ACM TOG (Proc. of SIGGRAPH) 27* (2008), 1–10. 1, 2, 7

[Deb06] DEBEVEC P.: A median cut algorithm for light probe sampling. *ACM SIGGRAPH 2006 Courses* (2006). 7

[DHT*00] DEBEVEC P. E., HAWKINS T., TCHOU C., DUIKER H.-P., SAROKIN W., SAGAR M.: Acquiring the reflectance field of a human face. In *SIGGRAPH* (2000), pp. 145–156. 3

[ECJ*06] EINARSSON P., CHABERT C.-F., JONES A., MA W.-C., LAMOND B., IM HAWKINS, BOLAS M., SYLWAN S., DEBEVEC P.: Relighting human locomotion with flowed reflectance fields. In *Proc. EGSR* (2006), pp. 183–194. 2, 3

[FH04] FELZENSZWALB P. F., HUTTENLOCHER D. P.: Efficient graph-based image segmentation. *IJCV 59* (2004). 4

[GCHS05] GOLDMAN D., CURLESS B., HERTZMANN A., SEITZ S.: Shape and spatially-varying brdfs from photometric stereo. In *ICCV* (oct. 2005), vol. 1, pp. 341–348. 2

[Geo03] GEORGHIADES A. S.: Recovering 3-d shape and reflectance from a small number of photographs. In *Proc. EGSR* (2003), EGRW '03, Eurographics Association, pp. 230–240. 2

[GSA*09] GALL J., STOLL C., AGUIAR E., THEOBALT C., ROSENHAHN B., SEIDEL H.-P.: Motion capture using joint skeleton tracking and surface estimation. In *Proc. IEEE CVPR* (2009), pp. 1746–1753. 1, 2, 3, 9

[HFB*09] HABER T., FUCHS C., BEKAERT P., SEIDEL H.-P., GOESELE M., LENSCH H. P. A.: Relighting objects from image collections. In *Proc. IEEE CVPR* (2009), pp. 627–634. 3, 4

[HVB*07] HERNANDEZ C., VOGIATZIS G., BROSTOW G. J., STENGER B., CIPOLLA R.: Non-rigid photometric stereo with colored lights. In *Proc. ICCV* (2007), pp. 1–8. 2

[Kaj86] KAJIYA J. T.: The rendering equation. In *Proc. SIGGRAPH* (1986), ACM, pp. 143–150. 4

[Li11] LI C.: *An Efficient Algorithm for Total Variation Regularization with Applications to the Single Pixel Camera and Compressive Sensing*. BiblioBazaar, 2011. 5

[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 TOG 22*, 2 (2003), 234–257. 2, 6

[MBR*00] MATUSIK W., BUEHLER C., RASKAR R., GORTLER S. J., MCMILLAN L.: Image-based visual hulls. In *SIGGRAPH* (2000), pp. 369–374. 1, 2

[MPBM03] MATUSIK W., PFISTER H., BRAND M., MCMILLAN L.: A data-driven reflectance model. *ACM TOG 22*, 3 (2003), 759–769. 2

[MPN*02] MATUSIK W., PFISTER H., NGAN A., BEARDSLEY P. A., ZIEGLER R., MCMILLAN L.: Image-based 3d photography using opacity hulls. *ACM TOG 21*, 3 (2002), 427–437. 2

[NRH04] NG R., RAMAMOORTHI R., HANRAHAN P.: Triple product wavelet integrals for all-frequency relighting. *ACM TOG 23*, 3 (2004), 477–487. 3, 5

[RH01a] RAMAMOORTHI R., HANRAHAN P.: An efficient representation for irradiance environment maps. In *SIGGRAPH* (2001), pp. 497–500. 3

[RH01b] RAMAMOORTHI R., HANRAHAN P.: A signal-processing framework for inverse rendering. In *SIGGRAPH* (2001), ACM, pp. 117–128. 2, 3, 5

[SGdA*10] STOLL C., GALL J., DE AGUIAR E., THRUN S., THEOBALT C.: Video-based reconstruction of animatable human characters. *ACM TOG (Proc. SIGGRAPH Asia) 29* (2010), 139:1–139:10. 1

[SH07] STARCK J., HILTON A.: Surface capture for performance based animation. *IEEE Computer Graphics and Applications 27(3)* (2007), 21–31. 1, 2

[SWI97] SATO Y., WHEELER M. D., IKEUCHI K.: Object shape and reflectance modeling from observation. In *SIGGRAPH* (1997), pp. 379–387. 2

[SZS*08] SZELISKI R., ZABIH R., SCHARSTEIN D., VEKSLER O., KOLMOGOROV V., AGARWALA A., TAPPEN M., ROTHER C.: A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE PAMI 30* (June 2008), 1068–1080. 4

[TAL*07] THEOBALT C., AHMED N., LENSCH H. P. A., MAGNOR M. A., SEIDEL H.-P.: Seeing people in different light-joint shape, motion, and reflectance capture. *IEEE TVCG 13*, 4 (2007), 663–674. 2, 4

[VBMP08] VLASIC D., BARAN I., MATUSIK W., POPOVIĆ J.: Articulated mesh animation from multi-view silhouettes. *ACM TOG (Proc. SIGGRAPH '08)* (2008). 1, 2

[VPB*09] VLASIC D., PEERS P., BARAN I., DEBEVEC P., POPOVIC J., RUSINKIEWICZ S., MATUSIK W.: Dynamic shape capture using multi-view photometric stereo. *ACM TOG 28*, 5 (2009), 174. 2

[WGT*05] WENGER A., GARDNER A., TCHOU C., UNGER J., HAWKINS T., DEBEVEC P.: Performance relighting and reflectance transformation with time-multiplexed illumination. *ACM TOG 24*, 3 (July 2005), 756–764. 3

[WVL*11] WU C., VARANASI K., LIU Y., SEIDEL H.-P., THEOBALT C.: Shading-based dynamic shape refinement from multi-view video under general illumination. In *Proc. IEEE ICCV* (2011). 2, 3, 4, 8

[WVT12] WU C., VARANASI K., THEOBALT C.: Full body performance capture under uncontrolled and varying illumination: A shading-based approach. In *ECCV (4)'12* (2012), pp. 757–770. 2

[WWC*05] WASCHBÜSCH M., WÜRMLIN S., COTTING D., SADLO F., GROSS M.: Scalable 3D video of dynamic scenes. In *Proc. Pacific Graphics* (2005), pp. 629–638. 2

[WWMT11] WU C., WILBURN B., MATSUSHITA Y., THEOBALT C.: High-quality shape from multi-view stereo and shading under general illumination. In *Proc. IEEE CVPR* (2011), pp. 969–976. 2

[YDMH99] YU Y., DEBEVEC P., MALIK J., HAWKINS T.: Inverse global illumination: recovering reflectance models of real scenes from photographs. In *SIGGRAPH* (1999), pp. 215–224. 2, 6

[YM98] YU Y., MALIK J.: Recovering photometric properties of architectural scenes from photographs. In *Proc. SIGGRAPH* (1998), SIGGRAPH '98, ACM, pp. 207–217. 2

[YPS10] YOON K.-J., PRADOS E., STURM P.: Joint estimation of shape and reflectance using multiple images with known illumination conditions. *IJCV 86*, 2-3 (2010), 192–210. 2