Intrinsic Textures for Relightable Free-Viewpoint Video

James Imber[1], Jean-Yves Guillemaut[2] and Adrian Hilton[2]

[1] Imagination Technologies Ltd., Kings Langley, Hertfordshire, UK
james.imber@imgtec.com
[2] Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, Surrey, UK
j.guillemaut@surrey.ac.uk, a.hilton@surrey.ac.uk

Abstract. This paper presents an approach to estimate the intrinsic texture properties (albedo, shading, normal) of scenes from multiple view acquisition under unknown illumination conditions. We introduce the concept of *intrinsic textures*, which are pixel-resolution surface textures representing the intrinsic appearance parameters of a scene. Unlike previous video relighting methods, the approach does not assume regions of uniform albedo, which makes it applicable to richly textured scenes. We show that intrinsic image methods can be used to refine an initial, low-frequency shading estimate based on a global lighting reconstruction from an original texture and coarse scene geometry in order to resolve the inherent global ambiguity in shading. The method is applied to relighting of free-viewpoint rendering from multiple view video capture. This demonstrates relighting with reproduction of fine surface detail. Quantitative evaluation on synthetic models with textured appearance shows accurate estimation of intrinsic surface reflectance properties.

Keywords: Free-Viewpoint Video Rendering, Image-Based Rendering, Relighting, Intrinsic Images

1 Introduction

Free-viewpoint video rendering (FVVR) gives the user the freedom to choose the viewpoint from which to view a captured scene [1–3]. FVVR has been applied successfully in sports TV production [4, 5] and video conferencing [6] among other applications. In FVVR, video from several cameras is used to reconstruct scene geometry using Multiple-View Stereo (MVS), and appearance is reproduced by projectively texturing the scene with the original images [7].

Recently, FVVR research has shifted from straightforward reproduction of the original scene to extending FVVR functionality [8, 9] with the goal of adapting it to other applications. In particular, the ability to relight an actor's performance for seamless compositing into arbitrary real-world and computer-generated surroundings is highly desirable, and is termed Relightable Free-Viewpoint Video Rendering (RFVVR).

Convincing RFVVR requires estimation of the parameters of a bidirectional reflectance distribution function (BRDF) for each point on the surface of a mesh from the appearance. The final appearance of a scene is a function of multiple parameters, including albedo, surface normals and specularity, as well as scene lighting, making the estimation of these parameters ambiguous. In this paper we address the problem of extracting intrinsic textures under arbitrary uncontrolled lighting, which is poorly constrained and requires that scene lighting be inferred together with the scene appearance.

The problem of fitting parameters (usually albedo and shading) to each pixel of an image, for which scene geometry is not available, has been studied extensively as intrinsic image extraction. This paper combines principles from intrinsic image extraction with prior knowledge of the scene to resolve the global ambiguity between shading and albedo. RFVVR and intrinsic image extraction approach the same problem from two angles - the former is a top-down approach, with knowledge of scene structure at its disposal, whereas the latter is a bottom-up approach, which relies on local image structure to decompose into albedo and shading images. In short, the coarse geometry available in RFVVR can be leveraged to resolve the global ambiguity present in intrinsic image reconstruction methods.

The proposed method improves on previous RFVVR methods, which make heavy use of the fact that albedo in a scene is likely to be piecewise constant. The piecewise constant albedo assumption breaks down in the presence of multi-albedo regions (such as wood or patterned fabric), and any subsequent surface refinement or normal map extraction will be invalid for such regions. For this reason, we propose a two-stage coarse-to-fine optimisation approach for albedo and shading. We use a segmentation-based coarse albedo estimate to estimate the lighting for the scene, after which the segmentation is discarded and we resort to a surface-based bilateral filter technique to estimate per-pixel albedo for complex materials. Finally, a highly-detailed surface normal map is extracted using the refined albedo, shading and irradiance estimates.

2 Related Work

Our approach to relighting draws from recent contributions in RFVVR and intrinsic image extraction. The availability of underlying geometry, and multiple viewpoints of the same scene, can be used as a powerful aid to the extraction of intrinsic images, which in the context of RFVVR are referred to as *intrinsic textures* as they are intrinsic appearance properties over the surface manifold.

2.1 Free Viewpoint Video Rendering

A scene model is reconstructed using MVS, which is projectively textured from the camera viewpoints [10]. To reproduce view-dependent aspects of appearance, such as specularity, camera views are blended together at run-time depending on viewpoint [11]. In the case of near-Lambertian scenes with accurate stereo

reconstruction and camera calibration, a single texture per frame can be produced without sacrificing quality, as has been done for the results presented in this paper.

2.2 Intrinsic Image Extraction

The problem of estimating albedo and lighting from an image, without knowledge of geometry, has been extensively studied in computer vision as the problem of intrinsic image extraction [12–15]. The interaction of physical objects with light is governed by its intrinsic colours (albedo), specular properties, transmission properties and surface normals. Any image of a physical scene can be decomposed into intrinsic images corresponding to each material property.

No knowledge of global scene shape is available in these image-based techniques, and they invariably require additional constraints (or assumptions) to be introduced for good global solutions to be found. For example, Bousseau et al. [16] has a user interact with the system to guide the process, whereas Barron and Malik [17] use a set of shape and albedo priors based on general localised properties of natural images.

2.3 Material Properties from Multi-View Video

RFVVR requires the estimation of shape and reflectance properties comprising the scene [18, 19]. Once the underlying geometry and surface reflectance properties are known, arbitrary lighting conditions can be introduced in what is then a conventional computer graphics rendering pipeline. This can be expressed as estimating the parameters of a BRDF. Commonly-used BRDFs include the Lambertian (diffuse reflection only) and Phong (a physically inaccurate, but simple) reflection models. Throughout this paper, the Lambertian reflectance model is used.

In this work, we combine prior shape estimates from MVS with intrinsic image texture estimation to resolve the inherent global ambiguity. This replaces the assumption of piecewise constant albedo [12] which has often been used in RFVVR to constrain reflectance estimation. This assumption commonly fails in natural scenes with textured surface appearance such as patterned fabric.

Active Lighting Controlling capture and lighting conditions allows highly accurate models of albedo and lighting behaviour to be estimated, since it reduces the number of unknown parameters. These systems are termed *active illumination* or *light stage* [18, 20]. This requires dedicated equipment, calibrated light sources and calibrated cameras. Active lighting is less practical for dynamic scenes, since it greatly increases the complexity of capture techniques. Einarsson et al. [21] demonstrate high-quality image-based relightable free-viewpoint video using a complex active capture system with time-multiplexed lighting. High-speed synchronised illumination and cameras, and post-registration of the images are required for reconstruction of reflectance properties and shape.

Fixed Calibrated Lighting Passive techniques, which have a fixed lighting arrangement for the duration of capture, are better suited to the problem of dynamic scene relighting. Lensch et al. [22] introduce a robust method for the extraction of time-varying BRDF given a coarse geometric model of a real-world object. They propose extraction of surface normals as well as albedo in fitting the BRDF to give the illusion of high-frequency geometry. Ahmed et al. [19] use a similar technique for relighting of free-viewpoint video. They use calibrated point sources to iteratively refine surface normals and albedos given coarse scene geometry. A regularisation term is imposed on the surface normal to discourage poor fits to the reference data. To help resolve ambiguity, a clustering method based on the piecewise-constant albedo assumption is used.

Uncalibrated Lighting More recently, the radiance from irradiance problem for Lambertian scenes is solved using spherical harmonics (SH) up to the second order [23]. Wu et al. [24] perform mesh refinement against the original images as opposed to normal extraction. Assuming a Lambertian reflectance model, the authors construct segment-based albedo and radiance estimates for each frame of the sequence. Local occlusion is used to resolve the radiance-from-irradiance problem to high SH orders.

This approach is extended to non-Lambertian cases in the work of Li et al. [9]. After solving the Lambertian radiance-from-irradiance problem, specular regions are used to localise light sources. The Phong model is fitted to the appearance. The techniques of both Wu et al. and Li et al. are for application to dynamic scenes, and temporal priors based on results from other frames form an important part of their methods.

Our proposed intrinsic texture approach performs a full-resolution fit of albedo and surface normal to the original texture. By contrast, Wu et al. and Li et al. optimise over the surface based on piecewise-constant albedo, which gives lower resolution due to the inherent smoothing of the regularisation. Our method accurately estimates the irradiance map for isolated frames, meaning that temporal priors from multi-frame sequences are not required.

3 Overview

We want to find albedo and surface normal textures for a coarse MVS scene reconstruction which give plausible results when rendered under arbitrary lighting conditions. No prior knowledge of lighting is assumed; the scenes were captured under unknown lighting conditions. To solve this problem, we propose first estimating the global scene irradiance, then using this to initialise localised refinement of albedo and shading, before finally fitting surface normals. An overview of the pipeline is given in Figure 1.

To estimate scene lighting, we start by coarsely segmenting the mesh surface into regions of similar albedo, making use of the observation that albedo is often piecewise constant. Unlike previous methods, this initial segmentation does not have to be accurate, and we make no attempt to refine it. Using this preliminary

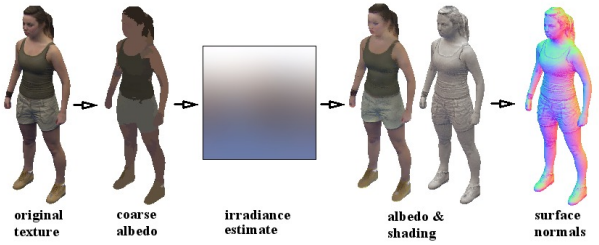

Fig. 1. Overview of the intrinsic texture extraction pipeline.

albedo estimate, we estimate the scene illumination which matches the shading distribution over the surface of the mesh. This provides a starting point for the albedo and shading texture extraction step, during which per-pixel albedo and shading textures are estimated. Finally, using the shading texture, surface normals are fitted to the lighting function. The normal map and albedo map can be used in conjunction to allow relighting of the FVV frame.

In using a coarse albedo estimate to determine the low-frequency global lighting, we do not lose any generality when applied to scenes with complex textures. The irradiance function is only recovered up to second order SH, meaning that any high-frequency variations in albedo within each segment will not corrupt the lighting estimate. Once the lighting has been estimated, the coarse albedo is discarded. This approach is in contrast to the current state-of-the-art method of first refining geometry based on a coarse albedo estimate, and then refining the BRDF parameters [9, 24]. It is thus capable of achieving full image-resolution albedo and surface normal maps for accurate surface detail.

4 Albedo and Shading Textures

The projectively textured, coarse MVS geometry is used to estimate the low-frequency irradiance. This irradiance accounts for the large, attached shadows at the scale of the MVS geometry, and is used to remove them from the original texture (section 4.1). To recover the missing high-frequency shading, an intrinsic image method is applied to the texture (section 4.2).

4.1 Low-Frequency Lighting Estimation

The global scene irradiance is reconstructed assuming Lambertian reflectance and infinitely displaced lighting. Ramamoorthi and Hanrahan [23] show that any irradiance map can be represented efficiently using spherical harmonics (SH) up to the second order, which requires only nine coefficients. This makes SH convenient for approximating the irradiance from a noisy set of samples.

The Lambertian reflectance model relates irradiance L to the radiance R by equation 1. The scene appearance I is related to the irradiance by $I(\boldsymbol{x}) =$

$A(\boldsymbol{x})L(\boldsymbol{x})$. $V(\theta, \phi, \boldsymbol{x})$ is a visibility mask which can only take the values 0 and 1.

$$L(\boldsymbol{n}(\boldsymbol{x}), \boldsymbol{x}) = \int_{\Omega} \max\left(\boldsymbol{u}(\theta, \phi)^{\intercal} \boldsymbol{n}(\boldsymbol{x}) \ , \ 0\right) R(\theta, \phi) \ V(\theta, \phi, \boldsymbol{x}) \ \mathrm{d}\Omega \tag{1}$$

$\boldsymbol{u}(\theta, \phi)$ is the unit vector in the direction of the spherical polar co-ordinates (θ, ϕ), and $\boldsymbol{n}(\boldsymbol{x})$ is the normal at surface position $\boldsymbol{x}$. The integral is over the sphere Ω with incremental surface area $\mathrm{d}\Omega = \sin(\theta)\mathrm{d}\theta\mathrm{d}\phi$. Under the assumption of a convex scene, the dependence on surface position $\boldsymbol{x}$ in equation 1 disappears, and this can be considered as a convolution of the radiance function with a large low-pass filter, termed the clamped-cosine kernel (equation 2).

$$L(\boldsymbol{n}) = \int_{\Omega} \max\left(\boldsymbol{u}(\theta, \phi)^{\intercal} \boldsymbol{n} \ , \ 0\right) R(\theta, \phi) \ \mathrm{d}\Omega \tag{2}$$

Due to this low-pass filtering, only spherical harmonics up to the second order can be reliably extracted in the case of convex objects [23]. Wu et al. extract the radiance function to higher orders by using the additional information provided by local self-occlusions in non-convex objects. For our purposes a low-order SH reconstruction of the irradiance suffices, since we rely on our intrinsic texture technique to extract high-frequency albedo, shading and surface normals. The lighting estimate is only used to globally balance the intrinsic albedo and shading textures in our case.

The texture is first segmented by albedo, using the segmentation of Felzenszwalb et al. [25] adapted to work in the tangent space of the mesh. This gives a set of materials, M. The material boundaries of this initial, coarse segmentation do not need to be pixel-accurate, since it is only used to recover the irradiance function. For each material u in M, an initial estimate of average albedo A'_u is given as the average colour of all texels (texture "pixels") comprising that material:

$$A'_u = \frac{1}{|u|} \sum_{\boldsymbol{x} \in u} I(\boldsymbol{x}) \quad \forall u \in M \tag{3}$$

In the case of monochrome lighting, this initial estimate of albedo is a scaled version of the final albedo, A_u, so that $k_u A_u = A'_u$. The problem of finding the correct ratios of material albedos A_u to each other is now a problem of determining the multipliers k_u.

The per-material coarse shading estimate is given by:

$$S_u(\boldsymbol{x}) = \frac{I(\boldsymbol{x})}{A'_u(\boldsymbol{x})} \tag{4}$$

Making use of the fact that the low-frequency shading can be considered as samples of the irradiance function, $S_u(\boldsymbol{x})$ can be projected along the coarse surface normal $\boldsymbol{n}_c(\boldsymbol{x})$ provided by the MVS scene reconstruction to give an estimate L'_u of the irradiance function at that point.

Fig. 2. Local irradiance estimates (left and centre) for two materials (polar projection). On the right, the intersection between the two irradiance estimates, Q_{ij}. Also shown are the positions of the materials on the mesh surface, highlighted in cyan.

$$L(\boldsymbol{n}_c(\boldsymbol{x})) \approx k_u L'_u(\boldsymbol{n}_c(\boldsymbol{x})) = k_u S_u(\boldsymbol{x}) \tag{5}$$

The sum of squared error in the overlap between the local irradiance estimates L'_u needs to be minimised by appropriate choices of k_u. For two materials $i, j \in M$, let $Q_{i,j}$ be the binary support function giving the overlap between L'_i and L'_j (Figure 2). The sum of squared error is given by:

$$E = \sum_i \sum_{j>i} \left[\int_\Omega (k_i L'_i(\theta,\phi) - k_j L'_j(\theta,\phi)) Q_{i,j}(\theta,\phi) \mathrm{d}\Omega \right]^2 \tag{6}$$

$$E = \sum_i \sum_{j>i} [k_i b_{ij} - k_j b_{ji}]^2 \tag{7}$$

$$\text{where} \quad b_{ij} = \int_\Omega L'_i(\theta,\phi) Q_{i,j}(\theta,\phi)\ \mathrm{d}\Omega \tag{8}$$

A greedy algorithm with a least-squares update step for each k_u is now used to minimise E. All k_u are initialised to 1. Since we are only interested in the ratios of the multipliers, the first multiplier, k_1, remains unchanged throughout, otherwise only the trivial solution $k_u = 0\ \forall u \in M$ would be found. $Q_{i,j}(\theta,\phi)$ does not take into account local occlusion of the lighting.

Let k_c represent the multiplier currently being optimised. Letting $d_j = k_j b_{ji}$, the update step is given by:

$$k_c \leftarrow \underset{k_c}{\operatorname{argmin}} ||k_c \boldsymbol{b}_c - \boldsymbol{d}||^2 = \frac{\boldsymbol{b}_c^\intercal \boldsymbol{d}}{\boldsymbol{b}_c^\intercal \boldsymbol{b}_c} \tag{9}$$

In this way, we can iterate over all the multipliers except for k_1, scaling the albedos to optimise the material overlaps, until convergence. This gives a refined estimate of the actual albedos A_u up to a global scale factor. These albedos may now be combined into a single coarse albedo estimate, A_c. The global irradiance estimate L is then found as the best fit of the SH basis up to the second order to L' (equation 10).

Fig. 3. Example intrinsic image decompositions (albedo and shading) using the proposed modified bilateral filter.

$$L'(\boldsymbol{n}_c(\boldsymbol{x})) = \frac{I(\boldsymbol{x})}{A_c(\boldsymbol{x})} \tag{10}$$

To test the effectiveness of this greedy approach, the order in which the materials were optimised was randomised. This was found to have no significant impact on the resulting $A_c(\boldsymbol{x})$. In the case of coloured irradiance, which is common in studio capture, the above can be done for each of the red, green and blue channels independently. It should be noted that this method only works on smooth meshes, since it relies on overlaps between per-material lighting estimates. In particular, it gives good results for human actors, but it would degrade for man-made objects with orthogonal faces.

4.2 Intrinsic Texture Extraction Filter

Our intrinsic texture extraction method builds upon the image-based method of Shen et al. [26] to incorporate global lighting information and operate over the surface of a mesh. To achieve this, a fast, bilateral filter based intrinsic image decomposition method is introduced. The use of an adaptive FIR filter for intrinsic image extraction, rather than explicitly minimising an energy functional, simplifies the method and is efficient in application to textures.

The contribution of Shen et al. [26] is an energy functional, which when minimised splits an image I into its constituent albedo A and shading S images, such that $I(\boldsymbol{x}) = A(\boldsymbol{x})S(\boldsymbol{x})$ (equation 11). It is show that this functional can be well approximated using a modified bilateral filter to remove local shading contributions from the original image.

$$E(A,S) = \sum_{\boldsymbol{x}\in P}\left(A(\boldsymbol{x}) - \sum_{\boldsymbol{y}\in N(\boldsymbol{x})} w(\boldsymbol{x},\boldsymbol{y})A(\boldsymbol{y})\right)^2 + \sum_{\boldsymbol{x}\in P}(I(\boldsymbol{x})/S(\boldsymbol{x}) - A(\boldsymbol{x}))^2 \tag{11}$$

$$w(\boldsymbol{x},\boldsymbol{y}) = \exp\left(-\frac{[\cos^{-1}(\hat{I}(\boldsymbol{x})^{\intercal}\hat{I}(\boldsymbol{y}))]^2}{\sigma_{i1}^2}\right)\exp\left(-\frac{[\text{luma}(I(\boldsymbol{x})) - \text{luma}(I(\boldsymbol{y}))]^2}{\sigma_{i2}^2}\right) \tag{12}$$

$$\text{luma} = 0.299 \times \text{Red} + 0.587 \times \text{Green} + 0.114 \times \text{Blue} \tag{13}$$

In equation 11, $N(\boldsymbol{x})$ is the neighbourhood of pixel $\boldsymbol{x}$, and P is the set of pixel positions. Equation 11 is made up of two parts. The first part imposes a metric for similarity in albedo between pixels which flattens out regions of similar albedo when minimised. The second part satisfies the condition that the observed image matches the estimated shading and albedo: $I(\boldsymbol{x}) = A(\boldsymbol{x})S(\boldsymbol{x})$. $S(\boldsymbol{x})$ is not dependent on the neighbourhood of pixels except through A, so a similar result can be achieved by minimising the following:

$$\underset{A}{\text{argmin}}\ E(A) = \sum_{\boldsymbol{x}\in P}\left(A(\boldsymbol{x}) - \sum_{\boldsymbol{y}\in N(\boldsymbol{x})} w(\boldsymbol{x},\boldsymbol{y})A(\boldsymbol{y})\right)^2 \tag{14}$$

Where $S = I/A$. This is equivalent to flattening out regions which are similar according to the metric defined in equation 12. This can be performed efficiently using a modified bilateral filter [27]:

$$A(\boldsymbol{x}) = \frac{1}{u}\int_{\boldsymbol{\mu}} I(\boldsymbol{\mu})\exp\left(-\frac{||\boldsymbol{x}-\boldsymbol{\mu}||_2^2}{\sigma_w^2}\right)\exp\left(-\frac{[\cos^{-1}(\hat{I}(\boldsymbol{x})^{\intercal}\hat{I}(\boldsymbol{\mu}))]^2}{\sigma_{i1}^2}\right) \times \exp\left(-\frac{[\text{luma}(I(\boldsymbol{x})) - \text{luma}(I(\boldsymbol{\mu}))]^2}{\sigma_{i2}^2}\right)\boldsymbol{d\mu} \tag{15}$$

In addition to the usual bilateral filter term which gauges similarity between pixels by luma, the chromaticity similarity term from equation 12 is also present. Some examples of image decompositions using this method are given in figure 3. The variances σ_{i1}^2 and σ_{i2}^2 adapt to the local region, as described in Shen et al.'s paper. u is a normalisation term to ensure the filter weights sum to unity.

The method of Shen et al. is based upon a local similarity metric, so it requires additional high-level constraints in order to achieve a good global solution. The same is true of the bilateral filtering based method described here. In the original paper, these constraints are provided by a user via a stroke-based interaction method, whereas we use the irradiance estimate from the MVS shape reconstruction to provide automatic global albedo balancing.

The quality of the results depends on the choice of kernel size in equation 15. Large kernels will have high variances σ_{i1}^2 and σ_{i2}^2, which will cause bleeding between regions which have similar albedo. In addition, it will take a long time to convolve large kernels with the image. Conversely, small kernels will not pick up large shading gradients, even with a large number of iterations. Throughout this paper, a 15x15 kernel was used.

Iteratively applying the kernel in equation 15 can reduce the computational cost and bleeding effects of using large kernels whilst still allowing some global shading effects to be extracted. The shading image is formed as $S = I/A$ after every iteration. The colour component of A should be preserved between

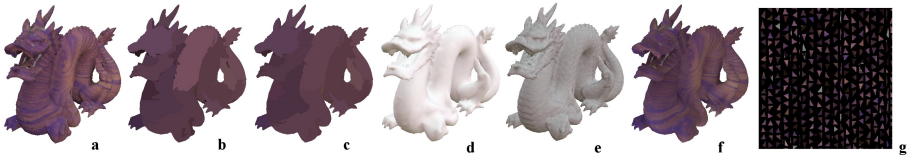

Fig. 4. Processing of original texture (a) to produce shading (e) and albedo (f) textures. Coarse albedos (b) are rebalanced (c) to allow a global shading estimate (d), which initialises the fine shading/albedo extraction method (equation 16). (g) shows part of the original texture in texture space.

iterations, which is equivalent to enforcing $\hat{A} = \hat{I}$ whilst preserving the RGB "length" of each pixel of A. We found only a single iteration to be necessary for the intrinsic texture results presented here.

This filter is adapted to work in the tangent space of the mesh by filtering directly on the texture in texture space. Where sample points fall off the edge of the triangle containing the centre of the filter, the sample point is offset to the triangle containing the required texel. To prevent distortion from mapping the surface onto the UV plane, the UV chart is split into individual triangles (Figure 4g). This preserves the shape and relative size of each triangle between the mesh and texture space. Figure 4 shows the result of albedo refinement on a texture. It was found that increasing the luma variance σ_{i2}^2 gives better results in the case of texture filtering.

To account for the global scene lighting, the original image is first divided by the irradiance estimate sampled using the coarse surface normals $\boldsymbol{n}_c(\boldsymbol{x})$:

$$W(\boldsymbol{x}) = \frac{I(\boldsymbol{x})}{L(\boldsymbol{n}_c(\boldsymbol{x}))} \tag{16}$$

The filter is then applied to W to obtain the albedo estimate A. The shading texture S is formed from I and A.

5 Refined Surface Normal Estimation

Since both a global lighting estimate and a surface shading estimate are available, it is possible to fit surface normals to the data. This is done by minimising an error function defined against the shading texture, S, at each point on the surface of the mesh $\boldsymbol{x}$:

$$E\big(\boldsymbol{n}(\boldsymbol{x})\big) = \big\| S(\boldsymbol{x}) - L\left(\boldsymbol{n}\right) \big\|_1 + \Lambda\left(\boldsymbol{n}, \boldsymbol{n}_c\right) \tag{17}$$

$$\boldsymbol{n}_{opt}(\boldsymbol{x}) = \underset{\boldsymbol{n}}{\operatorname{argmin}}\, E\big(\boldsymbol{n}(\boldsymbol{x})\big) \tag{18}$$

The L1 norm was chosen for its robustness in the presence of noise. When fitting surface normals, the MVS reconstruction gives a good indication of likely normal fits. Large deviations of the fitted normals $\boldsymbol{n}$ from the coarse normals

Table 1. Quantitative evaluation on synthetic datasets

Model	Shading Acc.	Colour Angle	Irradiance MSE	Time Taken
Smooth Sphere	0.911	3.162°	0.0059	102s
Rough Sphere	0.858	6.145°	0.0030	118s
Bunny	0.928	2.495°	0.0032	149s
Dragon	0.935	3.406°	0.0012	179s
Average	0.908	3.802°	0.0033	137s

$\boldsymbol{n}_c$ are unlikely, and are therefore penalised using a regularisation term Λ. To minimise this function, an exhaustive search of all possible fits in the direction of the gradient of the irradiance function is performed.

The regularisation term Λ is a function of the angle between the two vectors, defined as:

$$\Lambda\left(\boldsymbol{n}, \boldsymbol{n}_c\right) = \begin{cases} \lambda \left(\cos^{-1}(\boldsymbol{n}^\intercal \boldsymbol{n}_c)\right)^2 & \boldsymbol{n}^\intercal \boldsymbol{n}_c > 0 \\ \infty & \text{otherwise} \end{cases} \tag{19}$$

Where λ is determined experimentally. A value of 0.025 was used for all examples in this paper.

Since there is no inter-pixel dependency in equation 17, it is a good target for parallelisation. In our implementation, all surface normals are fitted in parallel on a GPU using a GLSL fragment shader. In all, the normal fitting stage takes of the order of 0.5 seconds to complete with a low-performance (Nvidia GeForce GT 240) graphics card, for a 1024x1024 texel texture.

6 Results

Ground-truth albedo and shading information is not available for multi-view sequences of actors. For this reason, a synthetic dataset consisting of multi-view renders of textured meshes, for which ground truth is available, is used to quantitatively evaluate the intrinsic texture method. Relit frames from public multiple view reconstruction datasets are also qualitatively evaluated to assess the performance for the target relightable FVVR application.

6.1 Quantitative Evaluation

A synthetic dataset was generated consisting of four models for evaluating the quality of the albedo and shading intrinsic images (Figure 5). The *bunny* and *dragon* models come from the Stanford 3D Scanning Repository. Each of these models was textured with a complex image, making albedo and lighting extraction comparatively difficult. A ray-tracer was used instead of a rasteriser to achieve more realistic lighting, including inter-reflections and ambient occlusions. A set of eight renders from virtual cameras arranged around the objects were combined into a single texture, in the same way as images from physical cameras

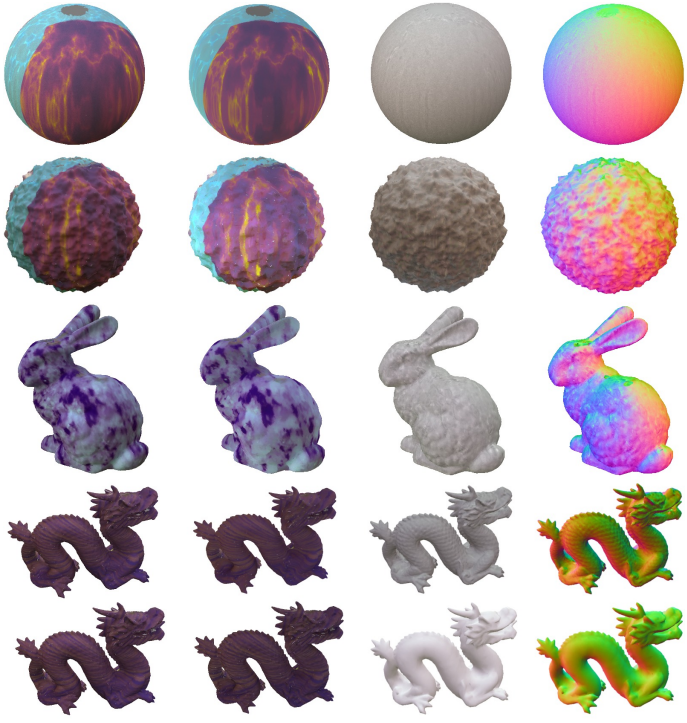

Fig. 5. Decomposition of synthetic textures and meshes into intrinsic textures. From left to right: original texture, albedo texture, shading texture and surface normals. The final row shows the result of omitting the local albedo/shading refinement stage.

are combined in the case of studio data. For each of the meshes, a low-polygon mesh (less than 6500 vertices in all cases) was used to generate the texture and provide coarse normals for irradiance and shading extraction. Ground truth shading and albedo was used for quantitative evaluation, with results given in Table 1. Our average runtime of 137 seconds per frame compares to 10 minutes reported by Wu et al. [24] and 7-8 minutes by Li et al. [9].

Two metrics were used to measure the accuracy of the separation between albedo and shading: shading accuracy and average colour angle, both measured in normalised RGB space (all axes in the range 0 to 1). Shading accuracy is given on a scale of 0 to 1 (equation 20).

$$\text{shading acc.} = 1 - \frac{1}{\sqrt{3}|P|} \sum_{\boldsymbol{x} \in P} ||A(\boldsymbol{x}) - A_G(\boldsymbol{x})||_2 \tag{20}$$

Here, P is the set of all texel positions, and $A_G(\boldsymbol{x})$ is the ground-truth albedo.

Fig. 6. Relighting of studio-captured data under general, uncalibrated lighting conditions using our method (see supplementary video). From left to right: original texture, albedo estimate, shading estimate and relighting under two different conditions. On far right, result from Li et al. [9]. Note the preservation of facial detail with the proposed method. Light spheres are shown on the top row.

The shading accuracy reflects the accuracy of the brightness of the albedo as well as R:G:B ratio. The colour angle is a measure of the accuracy of the R:G:B ratio only (equation 21).

$$\text{avg. RGB angle} = \frac{1}{|P|} \sum_{\boldsymbol{x} \in P} \cos^{-1} \left(\hat{A}(\boldsymbol{x})^{\intercal} \hat{A_G}(\boldsymbol{x}) \right) \tag{21}$$

A colour angle of zero indicates that the R:G:B ratios in the albedo estimate perfectly match those in the ground truth.

6.2 Qualitative Evaluation

The method was validated on three studio capture datasets and rendered under various lighting conditions. The first two sequences [8] were recorded with 8 cameras at a resolution of 1920x1080, whereas the last sequence [24] was recorded with 11 cameras at 1296x972. In all renders a high level of detail is achieved, and challenging textures are faithfully reconstructed, as shown in Figure **??**. In particular, the faces of the actors are reproduced accurately, which is vital for perceived realism. By using normal fitting rather than geometry refinement (Figure 7), we achieve a higher resolution in our relighting results than in current state-of-the-art methods [24, 9]. The supplementary video gives results for the full sequences.

The main shortcomings of this approach are misclassification of high-frequency dark albedos as shading, and noise in the extracted normal maps. The former is

Fig. 7. Surface normals before and after refinement.

most obvious as the representation of the edges of the dancer's t-shirt logo, and some facial features, in the shading texture. The noise in the relit images results because each texel in the extracted normal map is fitted independently of the others, so no neighbour-based smoothing takes place.

7 Conclusions and Future Work

This paper introduces a new method for reconstruction of accurate high resolution albedo and surface normal textures from approximate multiple view scene reconstruction with unknown illumination. The approach enables estimation of albedo, shading and surface normals at the resolution of the original texture. Unlike previous approaches, this approach does not assume regions of near-constant albedo, but also works with rich, multi-albedo textures. A novel bilateral filter approach is proposed for efficient shading refinement.

The proposed intrinsic texture estimation method is based on the observation that RFVVR and intrinsic image estimation are complementary approaches to appearance property estimation. It is shown that a global, low-frequency lighting estimate obtained from an original texture and coarse scene geometry can be used to initialise a local, high-frequency refinement step.

Intrinsic textures are applied to relighting of free-viewpoint rendering from multiple view video capture. This demonstrates relighting with reproduction of detailed surface appearance. Quantitative evaluation on synthetic models with non-uniform surface appearance shows accurate estimation of per-pixel albedo and normals.

A number of refinements to this method are possible. Improved global lighting estimation by solving the radiance-from-irradiance problem taking into account occlusions would give a more accurate global lighting estimate. Extension to non-Lambertian surfaces would improve the generality of the approach. Finally, additional temporal and spatial priors may further improve the quality of the intrinsic textures.

Acknowledgements. The authors would like to thank Imagination Technologies Limited for funding this research. We would also like to thank Chenglei Wu and Christian Theobalt for kindly providing access to their datasets.

References

1. Zitnick, C., Kang, S., Uyttendaele, M.: High-quality video view interpolation using a layered representation. ACM Transactions on Graphics **1**(212) (2004) 600–608
2. Matusik, W., Buehler, C., Raskar, R., Gortler, S.J., McMillan, L.: Image-based visual hulls. Proceedings of the 27th annual conference on Computer graphics and interactive techniques (2000) 369–374
3. Vedula, S., Baker, S., Kanade, T.: Image-based spatio-temporal modeling and view interpolation of dynamic events. ACM Transactions on Graphics **24**(2) (April 2005) 240–261
4. Kanade, T., Rander, P., Narayanan, P.: Virtualized reality: constructing virtual worlds from real scenes. IEEE Multimedia **4**(1) (1997) 34–47
5. Guillemaut, J.Y., Hilton, A.: Joint Multi-Layer Segmentation and Reconstruction for Free-Viewpoint Video Applications. International Journal of Computer Vision **93**(1) (December 2010) 73–100
6. Pan, C.H., Huang, S.C., Chang, Y.L., Lian, C.J., Chen, L.G.: Real-time free viewpoint rendering system for face-to-face video conference. Proceedings of IEEE International Conference on Consumer Electronics (2008) 1–2
7. Debevec, P., Taylor, C., Malik, J.: Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. Proceedings of the 23rd annual conference on Computer graphics and interactive techniques (1996) 1–10
8. Starck, J., Hilton, A.: Surface capture for performance-based animation. IEEE computer graphics and applications **27**(3) (2007) 21–31
9. Li, G., Wu, C., Stoll, C., Liu, Y., Varanasi, K., Dai, Q., Theobalt, C.: Capturing Relightable Human Performances under General Uncontrolled Illumination. Computer Graphics Forum **32**(2) (May 2013) 275–284
10. Debevec, P., Taylor, C., Malik, J.: Image-based modeling and rendering of architecture with interactive photogrammetry and view-dependent texture mapping. Circuits and Systems, Proceedings of the 1998 IEEE International Symposium on (1998) 14–17
11. Starck, J., Kilner, J., Hilton, A.: A Free-Viewpoint Video Renderer. Journal of Graphics, GPU, and Game Tools **14**(3) (January 2009) 57–72
12. Land, E.H., McCann, J.J.: Lightness and retinex theory. Journal of the Optical Society of America **61**(1) (January 1971) 1–11
13. Barrow, H., Tenenbaum, J.: Recovering intrinsic scene characteristics from images. Computer Vision Systems (1978) 3–26
14. Tappen, M.F., Freeman, W.T., Adelson, E.H.: Recovering intrinsic images from a single image. IEEE transactions on pattern analysis and machine intelligence **27**(9) (September 2005) 1459–72
15. Shen, L., Yeo, C.: Intrinsic images decomposition using a local and global sparse representation of reflectance. Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on (2011)
16. Bousseau, A., Paris, S., Durand, F.: User-assisted intrinsic images. ACM Transactions on Graphics **28**(5) (December 2009) 1
17. Barron, J., Malik, J.: Shape, albedo, and illumination from a single image of an unknown object. Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on (2012) 334–341
18. Debevec, P., Hawkins, T., Tchou, C.: Acquiring the reflectance field of a human face. Proceedings of the 27th annual conference on Computer graphics and interactive techniques (2000) 145–156

19. Ahmed, N., Theobalt, C., Seidel, H.p.: Spatio-temporal Reflectance Sharing for Relightable 3D Video. Computer Vision/Computer Graphics Collaboration Techniques **4418** (2007) 47–58
20. Matusik, W., Pfister, H., Ngan, A., Beardsley, P., Ziegler, R., McMillan, L.: Image-based 3D photography using opacity hulls. Proceedings of the 29th annual conference on Computer graphics and interactive techniques (2002) 427
21. Einarsson, P., Chabert, C., Jones, A., Ma, W.C., Lamond, B., Hawkins, T., Bolas, M., Sylwan, S., Debevec, P.: Relighting human locomotion with flowed reflectance fields. ACM Transactions on Graphics 2006 Sketches (2006)
22. Lensch, H.P.a., Kautz, J., Goesele, M., Heidrich, W., Seidel, H.P.: Image-based reconstruction of spatial appearance and geometric detail. ACM Transactions on Graphics **22**(2) (April 2003) 234–257
23. Ramamoorthi, R., Hanrahan, P.: On the relationship between radiance and irradiance: determining the illumination from images of a convex Lambertian object. Journal of the Optical Society of America A **18**(10) (2001) 2448
24. Wu, C., Varanasi, K., Liu, Y., Seidel, H.P., Theobalt, C.: Shading-based dynamic shape refinement from multi-view video under general illumination. 2011 International Conference on Computer Vision (November 2011) 1108–1115
25. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision **59**(2) (September 2004) 167–181
26. Shen, J., Yang, X., Jia, Y., Li, X.: Intrinsic images using optimization. Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on (2011)
27. Durand, F., Dorsey, J.: Fast bilateral filtering for the display of high-dynamic-range images. ACM Transactions on Graphics (TOG) (2002)