A Perceptually Validated Model for Surface Depth Hallucination

Figure 1: The leftmost and rightmost rendered images show synthetically relit textured surfaces based on albedo and surface depth acquired from a single view using diffuse-lit/flash-lit image pairs. The central (novel view) rendering was generated from a single image (shown in Figure 11(a)) obtained from an online texture resource and matched to a similar exemplar for which we have recovered albedo and depth using the image-pair method.

Abstract

Capturing depth to represent detailed surface geometry normally requires expensive specialized equipment and/or collection of a large amount of data. We introduce a practical, simple multiscale shape-from-shading method that takes diffuse-lit/flash-lit image pairs, and produces an albedo map and textured height field that can be viewed from any angle under any lighting. Using two lighting conditions enables us to subtract one from the other to estimate albedo. In the absence of a flash-lit image of a surface for which we already have a similar exemplar pair, we approximate albedo and diffuse shading using histogram matching. Our multiscale depth estimation is based on local visibility. Unlike other depth-from-shading approaches, all operations are performed in image space, and we impose no constant albedo restrictions. An experimental validation shows our method works for a broad range of textured surfaces, and users are frequently unable to identify our results as synthetic in a randomized presentation. Furthermore, they are unable to decide between a rendering of our depth map and an equivalent one generated from a laser range scan in side-by-side comparisons. We see this method as a significant advance in acquiring surface detail for texturing using a standard digital camera, with applications in architecture, archaeological reconstruction, games and special effects.

CR Categories: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Color, shading, shadowing, and texture;

Keywords: relighting, shape-from-shading, albedo estimation, textured surfaces, perception, computational photography

1 Introduction

Textured surfaces such as brick, stone, wood and many other building materials have local variations in their surface meso-structure. Shading variations due to self-shadowing provide important perceptual cues necessary to convey a correct impression of shape. Our objective is to produce synthetically relit results that are difficult to distinguish from photographs. (See Figure 1.) We show that an approximate representation of the real surface (depth + albedo map) may be used to relight a wide range of common textured surfaces in a visually plausible manner. To this aim, we introduce a practical method to recover approximate surface texture information from a single viewpoint. We take a 2D picture and infer depth where it isn’t fully divulged in the image so we call this depth hallucination.

Assessing the effect of new buildings on lighting requires detailed modeling of 3D geometry, materials reflectance characteristics or albedo (equivalent to diffuse reflectance) and knowledge of lighting conditions, so that simulations of appearance at different times of the day are possible. Our method is aimed primarily at the materials recovery part of such architectural reconstructions. However, this technique is equally applicable to recovering and representing surface detail for use in graphically rich games and movies. Currently, our method is being used to recover depth maps at a Mexican archaeological site [Blank for anonymity], for the production of a dome-projected movie.

Our main contribution is a novel, experimentally validated shape-from-shading method, which takes diffuse-lit/flash-lit image pairs and produces a plausible textured height field that can be viewed from any angle under any lighting. In the absence of a flash-lit image, we apply histogram matching against a visually similar texture for which we have recovered a model from diffuse-lit/flash-lit pairs. This practical optimization simplifies the capture requirements for large surfaces composed of the same material but containing significant meso-structure variation. To date, no published method for recovering and relighting textured height fields has been validated against equivalent photographs. Since our goal is to recover enough surface detail for plausible lighting, accuracy requirements are purely perceptual and are evaluated based on the final imagery. The experimental studies we conducted demonstrate that participants cannot reliably identify our relit images as synthetic, and more importantly that participants believe these to be as plausible as geometrically correct laser-scanned reconstructions.

2 Previous Work

Creating 3D models directly from photographs is appealing since it offers the potential for simple, fast, low-cost model acquisition
for photorealistic visualization [Debevec et al. 1996]. However, detailed surface meso-structure of building materials is rarely considered in such models. To correctly relight different materials requires separation of the way surfaces scatter light and the actual light striking the surface. Although solutions to separate these under specific constraints have been proposed [Narasimhan et al. 2003], the problem is not generally solvable. To fill in the missing information, humans use tacit knowledge gained from experience of real world illumination to estimate material properties and plausible lighting [Fleming et al. 2003]. A number of meso-structure recovery methods capture normal and texture maps with multiple sources [Rushmeier and Bernardini 1999; Lensch et al. 2003]. An accurate but data intensive approach is to capture and encode the appearance of textured surfaces with a gantry under a large number of lighting and viewing conditions [Dana et al. 1999]. A number of methods to recover albedo and meso-structure exist, but require sets of images and/or need specialized equipment [Yu et al. 1999; Liu et al. 2001; Li et al. 2006; Ngan and Durand 2006; Paterson et al. 2005; Paterson and Cain 2006].

Classic shape-from-shading solutions aim to acquire 3D depth information from a single image [Horn 1989; Malik and Maydan 1989; Leung and Malik 2001; Prados and Faugeras 2005]. This is an underconstrained problem. Numerous shapes, surface reflectances, and lighting conditions can give rise to the same shading pattern [Belghumeur et al. 1999], and associated ambiguities in shape perception [Ramachandran 1988; Langer and Bülthoff 2001]. However, shape-from-shading approaches are attractive for our application as they do not require special equipment or lengthy data-capture processes. Khan [2006] successfully demonstrated how, under certain circumstances, limitations in our ability to correctly interpret depth and lighting [Ostrovsky et al. 2005] can be exploited to create plausible synthetic images from a dark-is-deep approximation [Langer and Bülthoff 2000]. Our depth hallucination approach is inspired by their ideas.

A large body of literature on the topic of shape-from-shading exists, and we refer to published surveys for a review of existing methods [Zhang et al. 1999; Durou et al. 2007]. Broadly speaking, our approach performs irradiance estimation and is similar in spirit to the iterative technique of Langer and Zucker [1994]. Langer and Zucker’s model is specifically designed for recovering shape-from-shading on a cloudy day. They observe that when diffuse lighting, surface luminance depends primarily on a local aperture function, which is defined as the solid angle subtended by the visible sky at each surface point. They formulate a set of constraints, applying a robust numerical approach to solve for depth. However, there are a few practical limitations to their method. First, the model assumes uniform albedo, which is a problem for a wide range of textured surfaces. Second, their approach suffers from quantization since they perform discretized sampling of light source directions over a hemisphere at each point on a hypothetical surface. Third, their implementation is based on an iterative ray-tracing scheme, which is computationally expensive. Since our goal is to recover sufficient depth for plausibly relighting textured surfaces, we develop a simpler, deterministic image-space solution that approximates their results.

### 3 Depth Hallucination Method

Our process is illustrated in Figure 2. The individual steps are image capture, albedo and shading estimation, depth estimation, and relighting the surface. Specifics of how we estimate albedo and shading depend on whether the input to our process is a diffuse-lit/flash-lit image pair [Eisemann and Durand 2004], or a single diffuse-lit image. Subtracting the diffuse-lit image from the flash-lit image gives a reasonable estimate of albedo, and a comparison of our diffuse-lit image and albedo provides a usable estimate of diffuse shading for depth estimation. We discuss this in further detail in Section 3.2. Our depth estimation method is described in 3.3, and rendering of our final images is described in 3.4. Throughout these sections, we illustrate the steps in our process with a case study of a brick path and show the output of each intermediate step.

### 3.1 Image Capture

To capture our input images, we employ a standard digital SLR camera mounted on a tripod, and an attached strobe. If the textured surface contains significant specularities, cross-polarization (i.e., the polarizer on the flash is perpendicular to the polarizer on the lens) can be used to minimize highlights [Hershberger 2008].

First we capture a RAW format image\(^1\) under indirect illumination (i.e., overcast skies or shadow). We call this the diffuse-lit condition. A second photo is taken from the same point with the flash fired at full power. The camera is set to its maximum flash synchronization speed, while position, aperture, and focus are fixed to ensure good pixel registration between the diffuse-lit and flash-lit conditions. Ideally, the flash should be mounted as close to the camera lens as possible in order to minimize shadows, though the images shown in this paper were all taken with a standard flash mount. See Figure 3 for an example input image pair.

---

\(^1\)We use RAW format to simplify calibrating the images to each other, however our technique also works with JPEG images if they can be linearized.

---

Figure 2: Flow chart showing the steps in our process.

Figure 3: An example input photograph pair.
3.2 Albedo Map and Shading Image

The first stage in our method requires estimation of albedo and diffuse shading. We begin by calibrating our RAW image captures to one another based on their aperture $A$ (f-stop), ISO $I$, and shutter speed $T_s$ and convert to linear, floating-point pixel values using the following exposure correction factor ($C_e$):

$$C_e = \frac{A^2}{T_sI}$$ (1)

If absolute values were required, there would be an additional conversion factor which is unnecessary for relative measurements such as ours.

To calculate albedo $I_a(j)$ we perform the operation expressed below at each pixel $j$:

$$I_a(j) = \frac{I_f(j) - I_d(j)}{I_c(j)}$$ (2)

Pixel values in the diffuse-lit image $I_d$ are subtracted from our flash-lit capture $I_f$, and we divide the result by pixel values in the flash calibration image $I_c$ taken of a white Lambertian surface at a similar distance to correct for vignetting. This yields approximate reflectance values at each pixel. We apply a daylight white balance that provides a good match to the flash, therefore image subtraction results in a good color balance in our albedo image, as shown in Figure 4(a). In cases where flash shadows are present, we also apply a simple thresholding and neighbor-filling technique that copies detail from the flash-lit areas [Petschnigg et al. 2004]. In more severe cases, we can apply an intelligent shadow removal algorithm [Finlayson et al. 2006], though this requires some user intervention. (All examples shown in the paper used the simpler, automatic method.)

To compute the diffuse shading image, we take the ratio of the diffuse-lit condition over the albedo at each pixel. This can result in a color cast due to skylight or cloudy illumination, but our depth estimation method uses only the luminance channel. A grayscale computed shading image for our brick path is shown in Figure 4.

3.3 Depth Estimation

The Langer and Zucker method is designed to recover shape from shading on a cloudy day [Langer and Zucker 1994], which is precisely what we capture in our technique. Applying their relaxation method entails iteratively ray-tracing a discretely sampled hemisphere of light source directions at every surface point. Instead we develop an approximate solution that works entirely in image space and yields a direct estimate of depth at each pixel. A conservative set of assumptions ensure that we do not exaggerate depth variations, and a final user-adjusted scale factor is used to achieve the desired roughness.

Surface meso-structure can be approximated as a terrain with hills and valleys. The orientation of the surface to the sky (cosine factor) dominates on the hills, while the visible aperture effect dominates in the valleys, where the sides are at least partly in shadow. We therefore begin by developing two local models to approximate these two relationships between meso-structure depth and shading. The scope of each model is shown on a hypothetical textured surface in Figure 6. These models are derived such that an above-plane linear model is matched to a below-plane quadratic model, at a tangent point, creating the smooth piecewise function plotted in Figure 7.

Below-Plane Model

We derive our below-plane shadowing model by approximating pits in the surface as cylinders with an aperture $2a$ and depth $d$, as shown in Figure 8(a). In order to arrive at a simple formula, we chose to ignore interreflections which we found affect the scale but not the character of the depth estimates. We calculate an illumination factor $E_c$ by integrating the cosine weighting over the solid angle subtended by the visible sky:

$$E_c = \int \cos \theta \, dA$$

Figure 5: Example input image for histogram matching and generated shading image.
Through simple trigonometry the integrated shading factor becomes:

\[ S = \frac{E_c}{E_h} = \frac{\pi \sin^2 \theta}{\pi} = \frac{a^2}{a^2 + d^2} \] (4)

Pit depth can therefore be estimated by solving Equation 4 for \( d \) as:

\[ d = a \sqrt{\frac{1}{S} - 1} \] (5)

**Above-Plane Model**

For the above-plane model, we approximate surface protrusions as hemispheres whose shading is a function of the visible portion of the hemisphere \( h_v \) subtended by the solid angle \( \psi \) (Figure 8(b)) and added to the remaining reflected portion of the hemisphere outside the solid angle subtended by \( \psi, h_r \). Depth can thus be estimated by a simple linear model derived as follows, where \( \rho \) is the effective surrounding surface reflectance.

\[ h_v = \frac{\pi}{2} (1 + \cos \psi) \] (6)

\[ h_r = \frac{\rho \pi}{2} (1 - \cos \psi) \] (7)

Consequently our above-plane shading factor is calculated as the ratio of these quantities and \( \pi \):

\[ S = \frac{\frac{\pi}{2} (1 + \cos \psi) + \frac{\rho \pi}{2} (1 - \cos \psi)}{\pi} \] (8)

This can be simplified and solved for \( \cos \psi \) to give:

\[ \cos \psi = \frac{2S - (1 + \rho)}{1 - \rho} \] (9)

From Figure 8(b):

\[ d = R - R \cos \psi \] (10)

Substituting \( \cos \psi \) gives the linear model:

\[ d = 2R \frac{1 - S}{1 - \rho} \] (11)

where \( R \) is the radius of the hemispherical hill.

**Combined Model**

These two models, expressed in Equations 5 and 11 can be conveniently combined at a double root solution to their intersection, by substituting \( S = 1/2 \) and equating the corresponding values of \( d \) yielding:

\[ a = \frac{R}{1 - \rho} \] (12)

Again for convenience we assume the surrounding surface reflectance \( \rho = 0 \), allowing our modeled depth to be computed from the shading variation at each aperture scale, \( a \), from the combined formula:

\[ D(S) = \frac{d}{a} = \begin{cases} \sqrt{\frac{1}{S} - 1} & \text{for } S \leq 1/2 \\ \frac{2}{2(1 - S)} & \text{for } S > 1/2 \end{cases} \] (13)

**Multiscale Formulation**

In a multiscale formulation of our model, the value of \( a \) is taken from each blur radius in a Gaussian decomposition of the diffuse shading image, ensuring an appropriate scale is used for each shading feature. Starting from a normalized version of our diffuse shading image, as shown in Figure 4(b), we compute several Gaussian blurred images using kernel radii (\( r \)) increasing by powers of three up to a maximum detail size \( m \) based on image content and specified by the user. At each level, the image is divided by the image at the next largest kernel radius (upto the largest) and multiplied by 1/2 for normalization, effectively yielding a Laplacian pyramid of equal resolution images [Burt and Adelson 1983]. These blurred images are referred to as \( \ell(i) \). Each of our \( N \) levels is then transformed using the depth function \( D \) given in Equation 13, where \( a \) is replaced by the blur radius relative to the synthesized surface size at each pixel \( j \) arriving at a per pixel depth value \( D_j \):

\[ D_j = \sum_{i=1}^{N} \frac{r(i)}{m} [D(\ell_j(i)) - 1] \] (14)
Figure 9: The effect of different levels of Gaussian blur on the normalized shading image for the brick path example.

Figure 9 shows our progressively blurred shading images for the brick path example. We subtract 1 from our computed depths at each level since this is the normal value for $D(S)$ at the average image intensity of 0.5, and we want our average surface depth to be zero.

Figure 10: A comparison between a simple dark-is-deep approximation and our multi-scale model.

As noted earlier, our depth estimates are conservative. First, we ignored albedo to simplify our analysis. Second, in the interests of a simple combined model, we approximate indentations in the surface as pits, where a crevice model might be more appropriate in certain cases. We therefore apply a user-selected, uniform scaling factor to each depth map to compensate for this and achieve an acceptable visual match to the original surface appearance. For all our test scenes, this scaling factor was between 0.75 and 1.5. Our unoptimized implementation takes 15 seconds to generate a 900x600 resolution depth map on a single core 2.5 GHz desktop computer.

Figure 11: Depth map recovered from a single image of a rock wall obtained from the Web.

Figure 12: Results of relighting our brick path examples.

In Figure 10, we show the difference between our depth hallucination method and a global, linear, dark-is-deep approximation [Khan et al. 2006] applied to the same diffuse shading image. Notice that our model is less sensitive to noise and better approximates the upper surface as well as the crevices.

To demonstrate the flexibility of our approach, we downloaded a photograph of a textured rock wall from an online texture resource\(^2\) taken with an unknown camera, and recovered a depth map using histogram matching to a roughly similar surface to obtain a diffuse shading image. The original image and our hallucinated depth map are shown in Figure 11. A novel view of the same surface with a histogram-matched albedo image is shown in Figure 1 (center).

3.4 Relighting the Hallucinated Surface

Once we have an albedo map and a depth map for our surface, virtually any rendering algorithm may be applied. We use the Radiance physically-based renderer [Ward 1994] with a suitable sky model that includes both direct (solar) and indirect (sky) components, choosing a low angle of solar illumination to make our depth variations more visible. A directly lit surface will have a warmer color cast, since indirect outdoor (sky or overcast) illumination has a bluish tint, and we incorporate this in our model as well by adding a subtle warmth to the light color. Figure 12 shows the results for the brick path example with a solar altitude of 30°. Specularity is not specifically addressed in our method, but may be added trivially to the material model by assuming a uniform value, as might be encountered on a wet day. Our validation described in the following section addresses the visual plausibility of our rendered results.

4 Experimental Validation

We aim to answer two questions through two experiments. First, can our rendered images be reliably identified as synthetically generated? Second, do renderings generated using hallucinated depth maps appear plausible when compared with renderings using laserscanned data? We hypothesize that if users cannot reliably identify synthetically relit images created using our method while focused on assessing them, then we can conclude that our depth hallucina-

\[^2\]Source of image: http://www.texturewarehouse.com
The lighting frame was matched for specific image pairs (similar sequences. No in-filling techniques were applied to the laser-scan. physically-based rendering method was used for relighting both se-
to both the laser-scan and hallucinated depth maps, and the same depth map from photographs. The same albedo map was registered alent image frame generated using our technique for estimating the geometry acquired through a laser scanning process, and an equiv-
fixed viewpoint (but novel to the captured one) were used. Each animation depicting changing solar position over the scene with a
to ground-truth data. Twelve pairs of still image frames from an in which the aim was to evaluate our estimated depth maps relative
matched images. Due to the difficulty in acquiring photographs with natural sunny lighting conditions at exactly the same location, the set of equivalent day-lit photographs were not necessarily taken from an identical view point to the images used to recover texture hallucinations. Participants were asked to rank each image from 1 to 5, corresponding to their certainty that the image they were viewing was an untouched photograph. On this scale, we define 1 as definitely synthetically generated, and 5 as definitely an un-
touched photograph. The duration for which each image was dis-
played was determined via a pilot study involving 20 participants in which stimuli were presented for 1, 3 and 5 seconds. We found no significant differences in people’s ratings between images shown for given time intervals. Study data was collected for experiment one from a new set of 20 participants who were shown each stimu-
lus for 3 seconds.

4.2 Experiment Two

Our second study was a two-alternative forced-choice experiment in which the aim was to evaluate our estimated depth maps relative to ground-truth data. Twelve pairs of still image frames from an animation depicting changing solar position over the scene with a fixed viewpoint (but novel to the captured one) were used. Each image pair contained an image frame created using ground truth geometry acquired through a laser scanning process, and an equiv-
alent image frame generated using our technique for estimating the depth map from photographs. The same albedo map was registered to both the laser-scan and hallucinated depth maps, and the same physically-based rendering method was used for relighting both se-
quences. No in-filling techniques were applied to the laser-scan. The lighting frame was matched for specific image pairs (similar

5 Results and Data Analysis

In the first experiment, where participants rated how real the images looked, a repeated measures analysis of variance (ANOVA) showed a slight preference for the photographs (F\textsubscript{2.38} = 21.61, p<0.001). Unsurprisingly since reproducing the subtle effects captured in a photograph is a major challenge, this difference was statistically significant [TODO - effect size!]. An important result however is that on a scale of 1 to 5: photos received a mean rating of 3.97, relit images scored 3.22 for models derived with diffuse-lit/flash-

4.1 Experiment One

The goal of Experiment One was to assess whether people can reliably identify images created using our depth hallucination ap-
proach. Single images depicting a variety of textured surfaces, con-
sisting of both real photographs and synthetically relit images, were presented in a randomized order. (See Figure 1 and Figure 12 for examples.) A total of 27 images were presented to each person in this part of the study. This set contained 9 day-lit photographs, 9 synthetically relit images and 9 synthetically relit histogram-
matched images. Due to the difficulty in acquiring photographs with natural sunny lighting conditions at exactly the same location, the set of equivalent day-lit photographs were not necessarily taken from an identical view point to the images used to recover texture hallucinations. Participants were asked to rank each image from 1 to 5, corresponding to their certainty that the image they were viewing was an untouched photograph. On this scale, we define 1 as definitely synthetically generated, and 5 as definitely an un-
touched photograph. The duration for which each image was dis-
played was determined via a pilot study involving 20 participants in which stimuli were presented for 1, 3 and 5 seconds. We found no significant differences in people’s ratings between images shown for given time intervals. Study data was collected for experiment one from a new set of 20 participants who were shown each stimu-
lus for 3 seconds.
In the forced-choice experiment, a paired-sample t-test showed no significant difference between hallucinated depth and the laser-scan. This leads us to conclude that participants could not tell which of the two looked most plausible to them. Mean scores for each choice were 54% for the hallucinated depth, and 46% for the laser-scan. Nine participants out of 20 showed a preference for the hallucinated depth (see Figure 15), while 6 show a preference for the renderings based on the laser-scanned data. The remaining 5 seem undecided. The viewpoint for each data set was maintained identical across all stimuli and close to the captured view to avoid bias for or against either depth map. Within each image pair, the only variable was the depth map used to generate the image. If the view played a significant role in users’ assessments then the results would have been clearly bipolar, however only 2 participants consistently chose a particular depth map in every comparison. Between each image pair, lighting was varied. If lighting played a significant role in biasing the results, we would not have seen consistent strong differences in preference between subjects.

6 Limitations

Naturally, there are situations where our assumptions do not hold, and these may produce unexpected or undesired results. We examine two such cases, which we encountered while acquiring test scenes for our experiments.

The first case is shown in Figure 16(a), where ivy vines are physically separated from the stone surface below. The separation is small, but it violates one of our basic assumptions, which is that our surface may be plausibly represented as a height field. Even if our mathematical model held in this case, which it does not, our height field representation would still fail us. The vines appear to be protruding from the wall rather than next to it, and the rendering looks very wrong.

The second case is shown in Figure 16(b), where our surface is a reasonable match to our geometry assumptions, but the daylight illumination is not. In this area, the light comes primarily from one side, as it is nearby a dark structure and only a portion of the sky is visible on the cobbled ground. This results in a bias in the shading image, which our technique translates into a bias in the geometry, making the stones appear to lean towards the original sky direction. While this problem might be overcome with large bounce cards, but in a practical setting, such biases may be unavoidable, and would have to be corrected in a geometry post-processing step.

7 Conclusions

Ultimately, our goal is to combine hallucinated depth maps with geometry reconstructions of buildings. This is likely to pose further challenges, such as surfaces that are unevenly or directly lit, or where neighboring images are taken under differing illumination, or at irregular angles, and must be stitched together seamlessly. Our new multiscale shape-from-shading method for acquiring plausible surface textures, from single or paired captures, complements image-based reconstruction process by supplying surface detail.

In this work our objective was to accurately portray the altered appearance of textured surfaces under differing lighting and viewing conditions, while requiring only simple and practical data capture procedures. Starting from easily acquired diffuse-lit/flash-lit image pairs, we generate both an albedo map and textured height field, which can be relit and viewed from any angle under any lighting. Our model applies a surface aperture function but, in contrast to previous methods, works entirely in image space. If only a diffuse-lit image is available, we apply histogram matching with a similar exemplar pair with matching statistics to lift the flash requirement, further simplifying data capture. Compared to alternatives, such as laser scanning, our depth estimation method does not require special data registration, since both albedo and depth are acquired from perfectly aligned captures.

Histogram matching permits us to hallucinate local height variations from any diffuse-lit imagery, threfore gaps in our captured model may be filled-in using a texture synthesis technique [Efros and Freeman 2001]. Our processing from image to model is also sufficiently simple that depth and albedo maps could even be generated on the fly, from captured or synthetic texture data, on consumer-level graphics cards.

Experimental evaluation of this new approach yielded two important observations. First, although participants were able to distinguish photographs from synthetically relit images, they had real difficulty in doing this. Second, when presented with relit images 75% of participants rated them as more like photographs than synthesized images. Since depth is never fully divulged by shading, our depth estimates may fall short of absolute accuracy. However, participants were unable to decide whether hallucinated depth renderings or those generated using ground truth depth values acquired by laser scanning seemed most plausible.

References


