Shape, Albedo, and Illumination from a Single Image of an Unknown Object

Jonathan T. Barron and Jitendra Malik
UC Berkeley
{barron, malik}@eecs.berkeley.edu

Abstract

We address the problem of recovering shape, albedo, and illumination from a single grayscale image of an object, using shading as our primary cue. Because this problem is fundamentally underconstrained, we construct statistical models of albedo and shape, and define an optimization problem that searches for the most likely explanation of a single image. We present two priors on albedo which encourage local smoothness and global sparsity, and three priors on shape which encourage flatness, outward-facing orientation at the occluding contour, and local smoothness. We present an optimization technique for using these priors to recover shape, albedo, and a spherical harmonic model of illumination. Our model, which we call **SAIFS** *(shape, albedo, and illumination from shading) produces reasonable results on arbitrary grayscale images taken in the real world, and outperforms all previous grayscale "intrinsic image"-style algorithms on the MIT Intrinsic Images dataset.*

1. Introduction

We wish to take only a single grayscale image of an object and estimate the shape, albedo, and illumination that produced that image (Figure 1). This "inverse optics" problem is terribly underconstrained: the space of albedos, shapes, and illumination that reproduce an image is vast.

But of course, not all albedos and shapes are equally likely. Past work has demonstrated that simple statistics govern natural images [8, 23], and we will construct models of the similar statistics that can be found in natural albedo and shape. Our algorithm is simply an optimization problem in which we recover the most likely shape, albedo, and illumination under to our statistical model, such that a single image is exactly reproduced. Our priors are effective enough that shape, albedo, and illumination can be recovered from real-world images, and are general enough that they work across a variety of objects: a single model learned on teabags and squirrels can be applied to images of coffee cups and turtles. Our model can be seen in Figures 1, 2

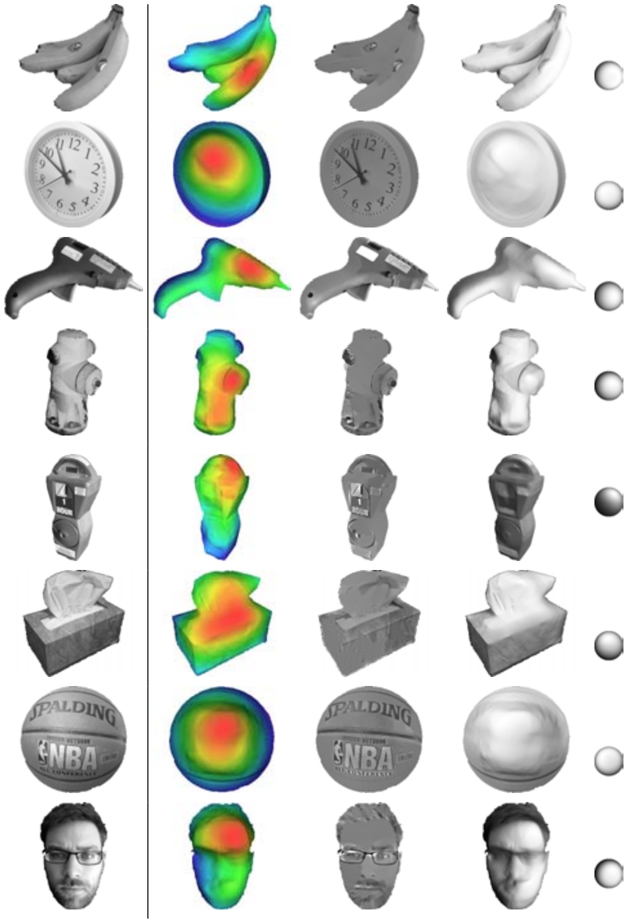


Figure 1. Our algorithm takes only a single masked grayscale image as input (shown on the left) and produces as output a depth map, albedo map, shading image, and spherical harmonic illumination model. These images were taken by the authors with a cellphone camera in uncontrolled indoor and outdoor illumination conditions. All images and results in this paper were produced using the same piece of code with the same parameter settings. The "shading" image is a rendering of the recovered depth under the recovered illumination. Depth is shown with a pseudo-color visualization (red is near, blue is far). Many more similar results can be found in the supplementary material.

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

Authorized licensed use limited to: University of Pennsylvania. Downloaded on January 12,2022 at 20:47:19 UTC from IEEE Xplore. Restrictions apply.

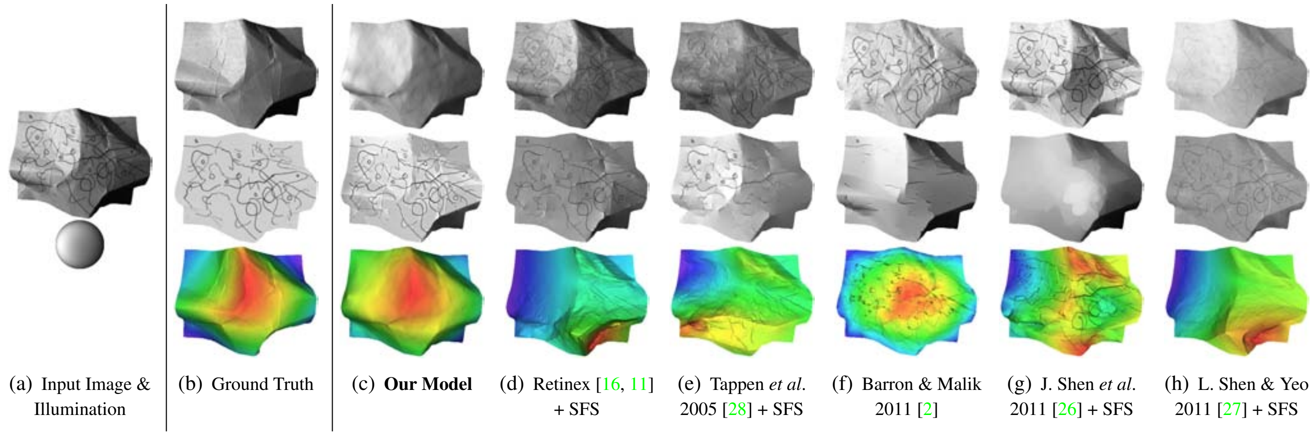

(a) Input Image & Illumination
(b) Ground Truth
(c) **Our Model**
(d) Retinex [16, 11] + SFS
(e) Tappen *et al.* 2005 [28] + SFS
(f) Barron & Malik 2011 [2]
(g) J. Shen *et al.* 2011 [26] + SFS
(h) L. Shen & Yeo 2011 [27] + SFS

Figure 2. A visualization of our technique, and others, for the task of recovering shading, albedo, and shape, given a single grayscale image and a known illumination. We benchmark against the other algorithms on the known-illumination task because no other algorithm we are aware of is capable of recovering shape, albedo, shading, and illumination for general objects.

and 7, and in the supplementary material.

The problem we address was first posed to the computer vision community as the "intrinsic images" problem [3], though it had been studied earlier in other fields as the problem of lightness constancy [10]. Over time the intrinsic images problem has been simplified to the problem of separating an image into shading and albedo [11, 16, 19], which has seen some recent progress [26, 27, 28] mostly by relying on color as a cue. However, none of these algorithms work well on grayscale images (see Figure 2, Table 1), nor do they allow shape (with the exception of [2]), or illumination to be directly recovered. Our model relates to the classic shape-from-shading problem [15] as we rely on shading as a cue for depth, but SFS techniques generally require albedo and illumination to be known, which we do not. Shape-from-contour is another classic single-image shape recovery technique [7, 18] which we build upon.

A competing approach to single-image techniques is to better constrain the problem with additional data. Instances of this approach are photometric stereo [30], structure from motion [12], and inverse global illumination [31]. All these techniques depend on multiple observations, and they all break down when given only a single image. Other work has explored recovering the spatial layout of a scene from a single image [14, 25], but these techniques do not recover albedo or illumination, and the recovered shape is very coarse. Morphable models [6] recover shape, albedo, and illumination, but require extremely specific models of the objects being estimated, and therefore do not work for general objects.

We build heavily on the problem formulation of Barron and Malik [2], which addressed a very constrained case of our problem. Their model required that a low-frequency observation of shape be known and that illumination be known, two fundamental limitations that prevent the technique from being useful "in the wild". This paper can be viewed as an extension to that work in which the priors on shape and albedo are dramatically improved, the requirement of low-frequency shape information is relaxed (due to the strength of our priors), and illumination is not required to be known.

Let us present a modification to the problem formulation of [2]. Assuming Lambertian reflectance and orthographic projection, given a log-albedo map A, a depth-map Z, and illumination L, the log-intensity image I is defined as $I = A + S(Z, L)$. The function $S(Z, L)$ is the "log-shading image" of Z with respect to L: it linearizes Z into a set of normals and renders those normals using L, a model of spherical harmonic illumination (detailed in the supplementary material). Now let us assume that the image I and illumination L have been observed, but Z and A are unknown. This problem is underconstrained, so we impose priors on Z and A and search for the most likely shape and albedo that explain image I. This is the same as minimizing the sum of $g(A)$ and $f(Z)$, which will be defined later as being (loosely) equivalent to the negative log-likelihoods of Z and A respectively. The optimization problem is:

$$\begin{aligned} \underset{Z,A}{\text{minimize}} \quad & g(A) + f(Z) \\ \text{subject to} \quad & I = A + S(Z, L) \end{aligned} \tag{1}$$

This is the same "shape and albedo from shading" (SAFS) formulation as in [2], but with different priors, a different rendering engine, and formulated in log-intensity space.

We will present $g(A)$, two priors[1] on albedo: one which encourages piecewise-smoothness by placing heavy-tailed

[1] Throughout this paper we use the term "prior" loosely. We refer to loss functions or regularizers on Z or A as "priors" because they have an interpretation as the unnormalized negative log-likelihood of some statistical model. We refer to minimizing entropy as a "prior", which is again an abuse of terminology. Our occluding contour "prior" requires first observing the silhouette of an object, and is therefore a posterior, not a prior.

Authorized licensed use limited to: University of Pennsylvania. Downloaded on January 12,2022 at 20:47:19 UTC from IEEE Xplore. Restrictions apply.

distributions on the multiscale gradient norm of log-albedo, and one which encourages a low entropy in the marginal distribution of log-albedo across all scales of an entire image. We then present $f(Z)$, three priors on shape: one which encourages fronto-parallel flatness (primarily to address the bas-relief ambiguity [4]), one which informs the surface orientation near an object's occluding contour, and one which encourages a novel measure of smoothness by placing heavy-tailed distributions on the multiscale gradient norm of the mean curvature of shape. We then extend the problem formulation in Equation 1 to present a novel framework for recovering illumination in addition to shape and albedo, using only our priors.

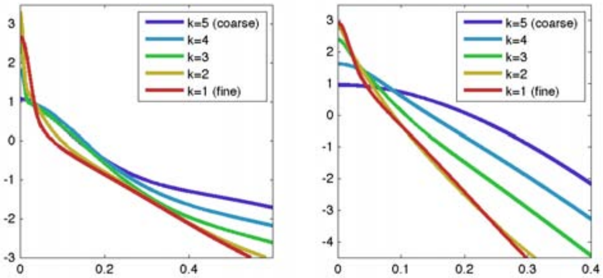
2. Priors on Albedo

We present two priors on albedo, one which encourages local piece-wise smoothness, and one which encourages albedo across the image to be composed (approximately) of a small number of values.

Our priors are on the differences of log-albedo, which makes them equivalent to priors on the ratios of albedo. This makes intuitive sense, as albedo is defined as a ratio of reflected light to incident light, but is also crucial to the success of our algorithm: Consider the albedo-map ρ implied by log-image I and log-shading $S(Z, L)$, such that $\rho = \exp(I - S(Z, L))$. If we were to manipulate Z or L to increase $S(Z, L)$ by some constant α across the entire image, then ρ would be divided by $\exp(\alpha)$ across the entire image, which would accordingly decrease the differences between pixels of ρ. Therefore, if we placed priors on the differences of albedo (as in [2]) it would therefore be possible to trivially satisfy our priors by manipulating shape or illumination to increase the intensity of the shading image. However, in the log-albedo case $A = I - S(Z, L)$, increasing all of S by α (increasing the brightness of the shading image) simply decreases all of A by α, and does not change the differences between log-albedo values. Priors on the differences of log-albedo are therefore invariant to scaling of illumination or shading, which means they behave similarly in well-lit regions as in shadowed regions, and cannot be trivially satisfied.

2.1. Smoothness

Albedo tends to be piecewise smooth — or equivalently, variation in albedo tends to be high-frequency and sparse. This is the insight that underlies the Retinex algorithm [11, 16, 19], and informs more recent intrinsic images work [2, 26, 27, 28].



(a) $\|\nabla A\|$ at different scales (b) $\|\nabla H(Z)\|$ at different scales

Figure 3. The log-likelihoods of Gaussian scale mixtures learned on $\|\nabla A\|$ and $\|\nabla H(Z)\|$ at different scales, using the training set of our dataset. We see similar heavy-tailed distributions as have been observed in natural image statistics [8, 23]. These distributions are the only learned parameters in our entire model.

Building on this idea, we use a simplification of the model of [2] (a multiscale Field-of-Experts-like model on albedo) and place heavy-tailed distributions on the gradient norm of log-albedo at multiple scales:

$$g_s(A) = \sum_{k=1}^{K} 4^{k-1} \sum_{x,y} c\left(\|\nabla \mathcal{G}(A,k)\|_{x,y} \,; \boldsymbol{\alpha}_A^k, \boldsymbol{\sigma}_A^k \right) \quad (2)$$

$\mathcal{G}(A, k)$ is the k-th level of a Gaussian pyramid of A, where $\mathcal{G}(A, 1) = A$. $\|\nabla \mathcal{G}(A,k)\|_{x,y}$ is the gradient norm of A at position (x, y) and level k. The 4^{k-1} multiplier accounts for the different number of pixels at each level of a Gaussian pyramid. K is equal to 5 in our experiments. In the supplementary material we detail how to calculate and differentiate $\|\nabla A\|$ efficiently using filter convolutions. $c(\,\cdot\,; \boldsymbol{\alpha}, \boldsymbol{\sigma})$ is the negative log-likelihood of a Gaussian Scale Mixture [21] parametrized by $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$, defined as:

$$c(x; \boldsymbol{\alpha}, \boldsymbol{\sigma}) = -\log \sum_{j=1}^{M} \alpha_j \cdot \mathcal{N}\left(x; 0, \sigma_j^2\right) \quad (3)$$

where $\boldsymbol{\alpha}$ are mixing coefficients, $\boldsymbol{\sigma}$ are the standard deviations of the Gaussians in the mixture, and M is the number of Gaussians ($M = 20$). The mean is set to 0, as the most likely albedo should be flat. The K GSMs are learned using EM on Gaussian pyramids of the albedo maps in our training set. The distributions we learn can be seen in Figure 3(a).

2.2. Minimal Entropy

Beyond piece-wise smoothness, the second property we expect from albedo is for there to be a small number of albedos in an image — that the albedo palette is small. As a hard constraint, this is not true: even in painted objects, there are small variations in albedo. But as a soft constraint, this minimum-entropy "sparsity" assumption holds. In Figure 4 we show the marginal distribution of log-albedo for three objects in our dataset. Though the man-made "cup"

Authorized licensed use limited to: University of Pennsylvania. Downloaded on January 12,2022 at 20:47:19 UTC from IEEE Xplore. Restrictions apply.

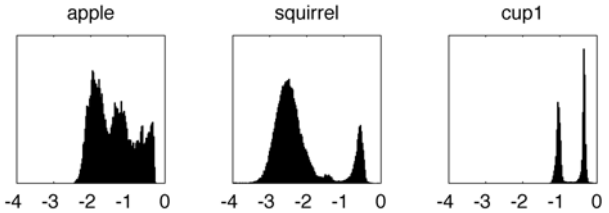


Figure 4. Three marginal distributions of log-albedo from our dataset. Log-albedo in an image tend to be grouped around certain values, or equivalently, these distributions tend to be low-entropy.

object shows the most clear peakedness in its distribution, the "natural" objects show significant clustering.

We are not the first to explore global "sparsity" priors on albedo. In [27] a sparsity constraint based on ℓ_1 minimization among global reflectance values was used. In [1], the assumption of albedo with a low Shannon entropy was used to resolve the bas-relief ambiguity [4]. We use the entropy formulation of [22], which has been previously used for the related task of shadow removal [9], to impose a similar prior on global albedo entropy. This entropy measure is defined as the negative-log of the following "information potential":

$$V(\mathbf{x}, \sigma) = \frac{1}{N^2\sqrt{4\pi\sigma^2}} \sum_{i=1}^{N} \sum_{j=1}^{N} \exp\left(-\frac{(x_i - x_j)^2}{4\sigma^2}\right) \quad (4)$$

In [22] this is derived as a measure of quadratic entropy (or Rényi entropy) for a set of points $\mathbf{x}$ assuming a Parzen window (a Gaussian kernel density estimator). Effectively, this is a "soft" and differentiable generalization of Shannon entropy, computed on a set of real values rather than a discrete histogram.

This entropy is quadratically expensive in N to compute naively. Others have used the Fast Gauss Transform (FGT) to approximate it in linear time [9], but the FGT does not provide a way to efficiently compute the analytical derivative of entropy. This was not a problem in [9], in which only two parameters were being optimized over, but our optimization involves gradient descent over entire albedo-maps, so we must be able to calculate the analytical derivative extremely efficiently.

This motivates our novel approximation to this measure of entropy. The key insight is that V can be re-expressed as:

$$V(\mathbf{n}, \sigma) \quad = \quad \mathbf{n}^{\mathrm{T}}(\mathbf{n} * \mathbf{g}) \quad (5)$$

where $\mathbf{n}$ is a histogram of $\mathbf{x}$, $*$ is convolution, and $\mathbf{g}$ is a particular Gaussian filter. Critically, this formulation also allows us to efficiently compute the gradient of $V(\mathbf{n}, \sigma)$ with respect to $\mathbf{n}$, which (provided $\mathbf{n}$ was generated using a "smooth" histogramming operation like linear interpolation) allows us to backpropagate the gradient onto $\mathbf{x}$. Our approximate entropy is extremely efficient to compute (usually $10\times$ to $100\times$ faster than using the FGT or the improved FGT) and is usually within 0.01% of the true entropy (similar to the accuracy of the FGT or IFGT). A thorough explanation of our technique is provided in the supplementary material.

The entropy of log-albedo A under this model is:

$$g_e(A) = -\sum_{k=1}^{K} \log\left(V(\mathcal{G}(A, k), \sigma_A)\right) \quad (6)$$

This is the sum of the entropies of each scale of a Gaussian pyramid of albedo. Modeling multiscale entropy in this fashion is moderately more effective than using a "flat" model of entropy on only the finest scale. Our final loss function on log-albedo is a linear combination of our smoothness and entropy terms:

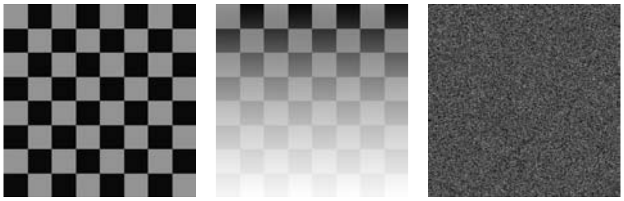
$$g(A) = \lambda_s g_s(A) + \lambda_e g_e(A) \quad (7)$$

where the λ multipliers and σ_A are learned through cross-validation on the training set.

At first glance, it may seem that our two priors are redundant: Encouraging piecewise smoothness seems like it should cause entropy to be minimized indirectly. This is often true, but there are common situations in which both of these priors are necessary. For example, if two regions are separated by a discontinuity in the image then optimizing for local smoothness will never cause the albedo on both sides of the discontinuity to be similar. The merit of both priors is further demonstrated in Figure 5.

3. Priors on Shape

Our prior on shape is three components: 1) a crude prior on flatness, to address the bas-relief ambiguity, 2) a prior



(a) $g_s = 23.2$, $g_e = \mathbf{7.9}$, $g_s + g_e = \mathbf{31.1}$

(b) $g_s = \mathbf{20.4}$, $g_e = 31.6$, $g_s + g_e = 52.0$

(c) $g_s = 42.4$, $g_e = \mathbf{7.9}$, $g_s + g_e = 50.3$

Figure 5. A demonstration of the importance of both components of our prior on albedo (only the finest scale of each prior is used). In 5(a) we have a natural looking albedo map, and the losses assigned by our prior on albedo. The other figures are shaded and shuffled versions of that same albedo. The shaded albedo has a lower smoothness cost than the original albedo, as its edges are less strong, but a higher entropy, as a larger variety of albedos are present. The shuffled albedo has the same entropy as the original, as its marginal distribution is identical, but has a much higher smoothness cost. Only the sum of the two costs correctly assigns the "natural" albedo map the lowest total cost.

Authorized licensed use limited to: University of Pennsylvania. Downloaded on January 12,2022 at 20:47:19 UTC from IEEE Xplore. Restrictions apply.

on the orientation of the surface normal near the occluding contour, and 3) a prior on smoothness in world coordinates, based on the variation of mean curvature.

3.1. Flatness

Because we are solving a superset of the shape-from-shading problem with unknown illumination, our priors must address the bas-relief ambiguity [4] (roughly, that absolute scale and orientation are ambiguous). We impose a prior that prefers the flattest shape within the bas-relief family, by minimizing the slant of Z at all points in Z:

$$f_f(Z) = -\sum_{x,y} \log\left(2N^z_{x,y}(Z)\right) \qquad (8)$$

Where $N^z_{x,y}(Z)$ is the z-component of the surface normal of Z at position (x, y) (as defined in the supplementary material), which increases as slant decreases. $f_f(Z)$ is the negative log-likelihood of the slant of Z assuming that the surface has been oriented uniformly at random in space. The particular form of $f_f(Z)$ is due to foreshortening: if we have observed a surface in space, it is more likely that it faces the observer ($N^z \approx 1$) than that it is perpendicular to the observer ($N^z \approx 0$).

3.2. Occluding Contours

The occluding contour of a shape is a powerful cue for shape interpretation [18], and algorithms have been presented for coarsely estimating shape given contour information [7]. At the occluding contour of an object, the surface is tangent to all rays from the vantage point. Under orthographic projection (which we assume), this means the z-component of the normal is 0, and the x and y components are determined by the contour in the image.

Because our dataset consists of masked objects, identifying the occluding contour C is trivial. For each point i on C, we estimate n_i, the local normal to the occluding contour in the image plane. Using those we regularize the surface normals in Z along the boundary by minimizing the following loss:

$$f_c(Z) = \sum_{i \in C} \sqrt{\left(N^x_i(Z) - n^x_i\right)^2 + \left(N^y_i(Z) - n^y_i\right)^2} \qquad (9)$$

Where $N(Z)$ is the surface normal of Z, as defined in the supplementary material. This loss function works better than more obvious alternatives, such as minimizing the angle between N_i and n_i. We believe this is due to the robustness of the square-root term (which allows the occluding contour assumption to be violated when necessary, see Figure 2) and because we do not directly minimize N^z (as N^z is very rarely close to 0 in our training set) but only implicitly minimize it by adjusting N^x and N^y.

3.3. Variation of Mean Curvature

There has been much work on modeling the statistics of natural shapes [2, 17, 29], with one overarching theme being that regularizing some function of the second derivatives of a surface is effective. However, this past work has severe issues with invariance to out-of-plane rotation and scale. Working within differential geometry, we present a representation of Z based on the variation of mean curvature, which allows us to place "smoothness" priors on Z that are invariant to rotation and scale.

To review: mean curvature is the divergence of the normal field. Planes and soap films have 0 mean curvature everywhere, spheres and cylinders have constant mean curvature everywhere, and the sphere has the smallest total mean curvature among all convex solids with a given surface area [13]. Mean curvature is a measure of curvature in *world coordinates*, not image coordinates, so (ignoring occlusion) the marginal distribution of $H(Z)$ is invariant to out-of-plane rotation of Z — a shape is just as likely viewed from one angle as from another. In comparison, the Laplacian of Z and the second partial derivatives of Z can be made large simply due to foreshortening, which means that priors placed on these quantities [2, 29] would prefer certain shapes simply due to the angle from which those shapes are observed — clearly undesirable.

But priors on just mean curvature are not scale-invariant. Were we to minimize $|H(Z)|$, then the most likely shape under our model would be a plane, while spheres would be unlikely. Were we to minimize $|H(Z) - \alpha|$ for some constant α, then the most likely shape under our model would be a sphere of a certain radius, but larger or smaller spheres, or a resized image of the same sphere, would be unlikely. Clearly, such scale sensitivity is an undesirable property for a general-purpose prior on natural shapes. Inspired by previous work on minimum variation surfaces [20], we place priors on $\|\nabla H(Z)\|$ — the change in mean curvature. The most likely shapes under such priors are surfaces of constant mean curvature, which are well-studied in geometry and include soap bubbles and spheres of any size (including planes). Priors on $\|\nabla H(Z)\|$, like priors on $H(Z)$, are invariant to rotation and viewpoint, as well as concave/convex inversion.

Mean curvature is defined as the average of principle curvatures: $H = \frac{1}{2}(\kappa_1 + \kappa_2)$. It can be approximated on a surface using filter convolutions that approximate first and second partial derivatives, as show in [5].

$$H(Z) = \frac{\left(1 + Z_x^2\right) Z_{yy} - 2 Z_x Z_y Z_{xy} + \left(1 + Z_y^2\right) Z_{xx}}{2\left(1 + Z_x^2 + Z_y^2\right)^{3/2}}$$

In the supplementary material we detail how to calculate and differentiate $H(Z)$ efficiently. We place heavy-tailed distributions on the gradient norm of the mean curvature of

Authorized licensed use limited to: University of Pennsylvania. Downloaded on January 12,2022 at 20:47:19 UTC from IEEE Xplore. Restrictions apply.

Z at multiple scales:

$$f_k(Z) = \sum_{k=1}^{K} 4^{k-1} \sum_{x,y} c\left(\left\| \nabla H \left(\frac{\mathcal{G}(Z,k)}{2^{k-1}} \right) \right\|_{x,y} ; \boldsymbol{\alpha}_Z^k, \boldsymbol{\sigma}_Z^k \right)$$

Notation is similar to Equation 2. The 2^{k-1} factor is necessary for shapes to be downsampled properly in z as well as x and y. $c(\,\cdot\,; \boldsymbol{\alpha}_Z^k, \boldsymbol{\sigma}_Z^k)$ is the negative log-likelihood of a GSM, parametrized by $\boldsymbol{\alpha}_Z^k$ and $\boldsymbol{\sigma}_Z^k$, which is trained on the k'th level of the Gaussian pyramids of the depth-maps in our training set. The distributions we learn for each scale can be seen in Figure 3(b).

Our final prior on shape is a linear combination of the three aforementioned components:

$$f(Z) = \lambda_f f_f(Z) + \lambda_c f_c(Z) + \lambda_k f_k(Z) \qquad (10)$$

where the λ multipliers are learned through cross-validation on the training set.

4. Optimization

We now address the problem of using our priors to recover shape and albedo given a known illumination. As in [2], we rearrange the problem in Equation 1 by defining A as a function of Z and I, reducing it to:

$$\underset{Z}{\text{minimize}} \qquad g(I - S(Z,L)) + f(Z) \qquad (11)$$

Because our priors are different, we can use more standard optimization techniques than in [2]. We optimize using a coarse-to-fine variant of L-BFGS: begining with a heavily downsampled Z, we optimize over Z using L-BFGS until convergence, then upsample Z by a factor of 2 and repeat optimization. Given a downsampled Z, calculation of the loss and the gradient of the loss is as follows: We upsample Z to the size of the image I, compute $S(Z,L)$ using our rendering engine, and compute the log-albedo A. We then compute $f(Z)$ and $g(A)$ and their gradients, and backpropagate $\nabla_A g(A)$ onto $\nabla_Z f(Z)$, which we then downsample to the size of the downsampled Z being optimized over. When optimization is complete, we have an estimated shape $\hat{Z}$, from which we compute an estimated shading $\hat{s} = \exp(S(\hat{Z},L))$ and albedo $\hat{\rho} = \exp(I - S(\hat{Z},L))$. For a single image, optimization to convergence takes between 1 and 10 minutes on a 2011 Macbook Pro, depending on the size of the input image. Backpropagating the analytical gradients of the loss functions can be somewhat daunting, so we provide implementation details in the supplementary material.

5. Unknown Illumination

To demonstrate how this technique can be modified to handle unknown illumination, we must reformulate our problem as one of MAP estimation:

$$\begin{aligned} \underset{Z}{\text{maximize}} \qquad & \log\left(P(A|Z,L)P(Z)\right) & (12)\\ P(A|Z,L) \quad = \quad & \exp(-g(I - S(Z,L)) & (13)\\ P(Z) \quad = \quad & \exp(-f(Z)) & (14) \end{aligned}$$

Until now we have assumed that L is known. However, we can introduce L as a latent variable, and marginalize over it:

$$\underset{Z}{\text{maximize}} \quad \log\left(P(Z) \sum_L P(A|Z,L)P(L) \right) \qquad (15)$$

Rather than optimize this marginal log-likelihood using EM (which is intractably expensive), or directly optimize the complete log-likelihood over Z and some L using gradient techniques (which works poorly due to local minima), we use the technique presented in [24] to directly optimize the expected complete log-likelihood using gradient-based techniques:

$$\underset{Z}{\text{maximize}} \quad \log\left(P(Z)\right) + \sum_L P(L|A,Z) \log\left(P(A|Z,L)\right) \qquad (16)$$

To optimize this we run our coarse-to-fine L-BFGS, but in each evaluation of the loss we compute the posterior over illumination $P(L|A,Z)$ (equivalent to the E-step), and use this to compute the marginal log-likelihood and its gradient. When L-BFGS takes a step, this is equivalent to a partial M-step. $P(L)$ is a simple model which assigns a uniform probability to all illumination conditions in our training set (some of which are shown in Figure 6). $P(L)$ and our optimization technique are detailed further in the supplementary material. Once optimization is complete, we use our recovered shape $\hat{Z}$ to compute the expected illumination $\hat{L} = \sum_L P(L|A,\hat{Z})L$, and use that to produce $\hat{s}$ and $\hat{\rho}$.

Naive optimization requires evaluating $P(A|Z,L)$ for $\sim$100 different illumination conditions each time we evaluate our loss function. In the supplementary material we present a simple approximation which makes optimization more tractable, but optimization is still $\sim$30× more expensive than when illumination is known. Results are shown in Table 1, Figures 1 and 7, and the supplementary material.

6. Results

We present an augmented version of the MIT Intrinsic Images dataset [11] in which we have used photometric stereo to estimate the shape of each object and the spherical harmonic illumination for each image. An example is shown in Figure 6, and details are provided in the supplementary material. In all of our experiments, we use the following test-set: cup2, deer, frog2, paper2, pear, potato, raccoon, sun, teabag1, turtle. The other 10 objects are used for training.

Authorized licensed use limited to: University of Pennsylvania. Downloaded on January 12,2022 at 20:47:19 UTC from IEEE Xplore. Restrictions apply.

Let us define some error metrics:

$$Z\text{-MAE} = \frac{1}{n} \min_{\beta} \sum_{x,y} \left| \hat{Z}_{x,y} - Z^*_{x,y} + \beta \right| \tag{17}$$

$$S\text{-MSE} = \frac{1}{n} \min_{\gamma} \sum_{x,y} \left(\gamma \hat{s}_{x,y} - s^*_{x,y} \right)^2 \tag{18}$$

$$\rho\text{-MSE} = \frac{1}{n} \min_{\alpha} \sum_{x,y} \left(\alpha \hat{\rho}_{x,y} - \rho^*_{x,y} \right)^2 \tag{19}$$

$$I\text{-MSE} = \frac{1}{n} \sum_{x,y} \left\| \hat{\alpha} \hat{\rho}_{x,y} \hat{N}_{x,y} - \rho^*_{x,y} N^*_{x,y} \right\|^2 \tag{20}$$

Z-MAE is the shift-invariant mean absolute error of $\hat{Z}$ (absolute error is used instead of squared error because it has an interpretation as the volume of the error). S-MSE and ρ-MSE are scale-invariant MSE of the recovered shading $\hat{s}$ and albedo $\hat{\rho}$. The scale invariance accounts for the ambiguity in the absolute brightness of the scene or absolute intensity of the albedo. I-MSE is the expected MSE in re-rendering $\hat{Z}$ and $\hat{\rho}$ under different illuminations, relative to the ground truth. This error metric was defined in [2], but due to an error in that paper's calculations, we reformulate it here (see the supplementary material for the derivation). I-MSE is made to be roughly scale invariant by using the scaled version of ρ we obtain when computing ρ-MSE. We also use LMSE, the locally scale-invariant error in shading and albedo defined in [11].

Results for recovering shape and albedo given a single image and a known illumination are in Table 1. We compare against several recently-published intrinsic images algorithms (meant to decompose an image into shading and albedo components), upon which we've run a shape-from-shading algorithm on the shading image. Though many of the intrinsic images algorithms were designed for color images, we've run them on grayscale images. For the sake of a generous comparison, the SFS algorithm uses our priors on Z, which boosts each baseline's performance (detailed in the supplementary material). We also compare against [2], which directly produces shape and albedo. The "flat" algorithm is a baseline in which $Z = 0$. Our performance relative to all other algorithms is visualized in Figure 2 and in the supplementary material.

Table 1 also includes an ablation study, in which components of our model are removed, and includes a shape-from-contour model in which only our priors on Z are used. The ablation study shows that removing any prior has a negative effect on performance. Albedo entropy has a relatively small impact on our error metrics, but produces noticeable improvements in the output's appearance.

Table 1 also shows our performance when illumination is unknown. Surprisingly, performance is very similar whether or not illumination is known. Our priors on shape and albedo (and our simple prior on illumination) appear to be sufficient to accurately recover illumination in addition to shape and albedo.

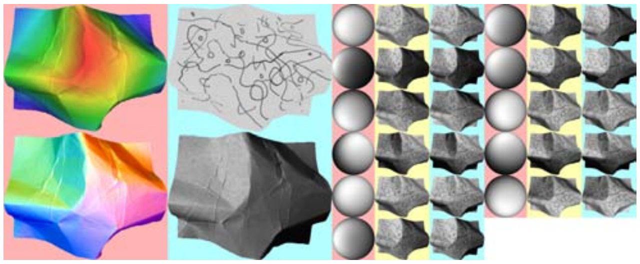


Figure 6. An object from our dataset. The MIT Intrinsic Images dataset's [11] contributions are shown in blue (ground-truth shading and reflectance, 10 images from different illuminations, and a "diffuse" image). We generate "ground truth" shape and illumination (and surface normals, implicitly), shown in red. The yellow images are renderings of our shape and illuminations, which show that our recovered shapes and illuminations are reasonable. The diffuse image (and optionally, its illumination) in the bottom row are the only input to our algorithm, and the other components are used for evaluating our results (if the object is in the test set) or for learning our priors (if the object is in the training set).

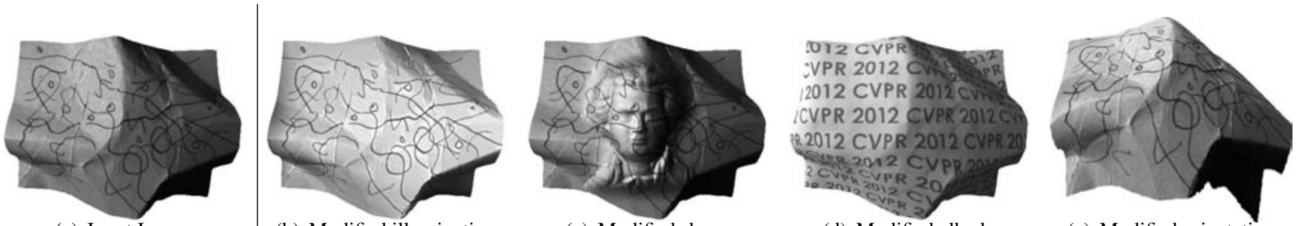
To clarify, all the images and results in this paper were produced using the same piece of code and (except those in the ablation study) the same setting of our 5 hyperparameters: $\lambda_s, \lambda_e, \sigma_A, \lambda_k, \lambda_f, \lambda_c$, which were all tuned on the training set. Performance could be improved by tuning these parameters for each image, which we did not do.

Recovering **Shape and Albedo** given Illumination and 1 Grayscale Image

Algorithm	Z-MAE	I-MSE	LMSE	S-MSE	ρ-MSE	Avg.
Flat Baseline	25.56	0.1369	0.0385	0.0563	0.0427	0.2004
Retinex [11, 16] + SFS	82.06	0.1795	0.0289	0.0291	0.0264	0.2009
Tappen *et al*. 2005 [28] + SFS	43.30	0.1522	0.0292	0.0343	0.0256	0.1761
Barron & Malik 2011 [2]	21.10	0.0829	0.0584	0.0282	0.0468	0.1682
J. Shen *et al*. 2011 [26] + SFS	48.51	0.1629	0.0445	0.0478	0.0450	0.2376
L. Shen & Yeo 2011 [27] + SFS	31.61	0.1191	0.0205	0.0236	0.0174	0.1260
Our Shape from Contour	21.42	0.0805	0.0350	0.0280	0.0311	0.1394
Our Model (No $\|\|\nabla A\|\|$)	17.50	0.0620	0.0289	0.0188	0.0238	0.1070
Our Model (No $\|\|\nabla H(Z)\|\|$)	21.81	0.1011	0.0341	0.0205	0.0194	0.1244
Our Model (No Flatness)	35.11	0.0651	0.0190	0.0148	0.0157	0.1002
Our Model (No Contour)	28.45	0.0811	0.0204	0.0167	0.0189	0.1082
Our Model (No Albedo Entropy)	21.23	0.0523	0.0196	0.0138	0.0162	0.0865
Our Model (All Priors)	21.86	0.0521	0.0191	0.0136	0.0156	0.0856
Recovering **Shape, Albedo, and Illumination** given 1 Grayscale Image						
Our Model (All Priors)	19.41	0.0577	0.0197	0.0178	0.0193	0.0946

Table 1. A comparison of our model against other intrinsic images algorithms. Shown are the five error metrics (the geometric mean across the test set), and an "average" error (the geometric mean of the other five geometric means, and what we minimize during cross-validation). Also shown are special cases of our model in an "ablation study", in which priors are removed. Each component contributes positively to the performance of our complete model. For the unknown illumination task, we were not able to find any algorithms to compare against. Note that the "L. Shen & Yeo 2011 [27]" baseline hand-tunes two parameters for each image in the test set to minimize LMSE, while we use a single set of parameters for all images, tuned to the training set.

Authorized licensed use limited to: University of Pennsylvania. Downloaded on January 12,2022 at 20:47:19 UTC from IEEE Xplore. Restrictions apply.



(a) Input Image (b) Modified illumination (c) Modified shape (d) Modified albedo (e) Modified orientation

Figure 7. Our system has obvious graphics applications. Given only a single image grayscale image, we can estimate an object's shape, albedo, and illumination, modify any of those three scene properties (or simply rotate the object), and then re-render the object.

7. Conclusion

We have presented a series of novel priors on shape and albedo. For albedo, we have explored smoothness and global sparsity in log-albedo space. For shape, we have explored flatness, the orientation of the surface at the occluding contour, and second-order smoothness using the variation of mean curvature. Previous work on jointly inferring shape and albedo was extended to allow for illumination to be estimated as well without a decrease in accuracy.

Shading is our algorithm's primary cue for inferring depth, and in a shading-free scene (in which all illumination is ambient) our model reduces to a shape-from-contour algorithm. Cast shadows and specularities are currently not addressed by our model, and are often the cause of errors. Our technique is targeted towards objects; to be extended to scenes, one could use segmentation techniques to generate candidate objects.

Our model produces qualitatively reasonable results on arbitrary grayscale images taken in the real world, and dramatically outperforms all previously published algorithms on the MIT Intrinsic Images dataset. Our technique, to the best of our knowledge, is the first unified model for jointly estimating shape, albedo, and illumination from a single image.

References

[1] N. Alldrin, S. Mallick, and D. Kriegman. Resolving the generalized bas-relief ambiguity by entropy minimization. *CVPR*, 2007. 4

[2] J. T. Barron and J. Malik. High-frequency shape and albedo from shading using natural image statistics. *CVPR*, 2011. 2, 3, 5, 6, 7

[3] H. Barrow and J. Tenenbaum. *Recovering Intrinsic Scene Characteristics from Images*. Academic Press, 1978. 2

[4] P. Belhumeur, D. Kriegman, and A. Yuille. The Bas-Relief Ambiguity. *IJCV*, 1999. 3, 4, 5

[5] P. Besl and R. Jain. Segmentation through variable-order surface fitting. *TPAMI*, 1988. 5

[6] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. *SIGGRAPH*, 1999. 2

[7] M. Brady and A. Yuille. An extremum principle for shape from contour. *TPAMI*, 1983. 2, 5

[8] D. Field. Relations between the statistics of natural images and the response properties of cortical cells. *JOSA A*, 1987. 1, 3

[9] G. D. Finlayson, M. S. Drew, and C. Lu. Entropy minimization for shadow removal. *IJCV*, 2009. 4

[10] A. Gilchrist. *Seeing in Black and White*. Oxford University Press, 2006. 2

[11] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman. Ground-truth dataset and baseline evaluations for intrinsic image algorithms. *ICCV*, 2009. 2, 3, 6, 7

[12] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge Univ Press, 2003. 2

[13] D. Hilbert and C. S. Vossen. *Geometry and the Imagination*. Chelsea Publishing Company, 1956. 5

[14] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 2007. 2

[15] B. K. P. Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. Technical report, MIT, 1970. 2

[16] B. K. P. Horn. Determining lightness from an image. *Computer Graphics and Image Processing*, 1974. 2, 3, 7

[17] J. Huang, A. B. Lee, and D. Mumford. Statistics of range images. *CVPR*, 2000. 5

[18] J. Koenderink. What does the occluding contour tell us about solid shape. *Perception*, 1984. 2, 5

[19] E. H. Land and J. J. McCann. Lightness and retinex theory. *JOSA*, 1971. 2, 3

[20] H. P. Moreton and C. H. Squin. Functional optimization for fair surface design. In *SIGGRAPH*, 1992. 5

[21] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli. Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Trans. Image Process*, 2003. 3

[22] J. C. Principe and D. Xu. Learning from examples with quadratic mutual information. *Workshop on Neural Networks for Signal Processing*, 1998. 4

[23] D. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 1994. 1, 3

[24] R. Salakhutdinov, S. Roweis, and Z. Ghahramani. Optimization with em and expectation-conjugate-gradient. *ICML*, 2003. 6

[25] A. Saxena, M. Sun, and A. Ng. Make3d: learning 3d scene structure from a single still image. *TPAMI*, 2008. 2

[26] J. Shen, X. Yang, Y. Jia, and X. Li. Intrinsic images using optimization. *CVPR*, 2011. 2, 3, 7

[27] L. Shen and C. Yeo. Intrinsic images decomposition using a local and global sparse representation of reflectance. *CVPR*, 2011. 2, 3, 4, 7

[28] M. F. Tappen, W. T. Freeman, and E. H. Adelson. Recovering intrinsic images from a single image. *TPAMI*, 2005. 2, 3, 7

[29] O. Woodford, P. Torr, I. Reid, and A. Fitzgibbon. Global stereo reconstruction under second-order smoothness priors. *TPAMI*, 2009. 5

[30] R. Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 1980. 2

[31] Y. Yu, P. Debevec, J. Malik, and T. Hawkins. Inverse global illumination: recovering reflectance models of real scenes from photographs. *SIGGRAPH*, 1999. 2

Authorized licensed use limited to: University of Pennsylvania. Downloaded on January 12,2022 at 20:47:19 UTC from IEEE Xplore. Restrictions apply.