Intrinsic Decomposition of Image Sequences from Local Temporal Variations

Pierre-Yves Laffont
ETH Zurich

Jean-Charles Bazin
ETH Zurich

Abstract

We present a method for intrinsic image decomposition, which aims to decompose images into reflectance and shading layers. Our input is a sequence of images with varying illumination acquired by a static camera, e.g. an indoor scene with a moving light source or an outdoor timelapse. We leverage the local color variations observed over time to infer constraints on the reflectance and solve the ill-posed image decomposition problem. In particular, we derive an adaptive local energy from the observations of each local neighborhood over time, and integrate distant pairwise constraints to enforce coherent decomposition across all surfaces with consistent shading changes. Our method is solely based on multiple observations of a Lambertian scene under varying illumination and does not require user interaction, scene geometry, or an explicit lighting model. We compare our results with several intrinsic decomposition methods on a number of synthetic and captured datasets.

1. Introduction

Intrinsic image decomposition aims to express an image of a Lambertian scene as the per-pixel product of reflectance and shading layers [3]. The reflectance component, also called albedo, represents how a diffuse surface reflects light (i.e., the material color), while the shading component represents the lighting effects such as shadows and indirect lighting. This decomposition is useful for several image editing applications such as recolorization [25], re-texturing [7], and relighting [20], among many others.

Intrinsic image decomposition is ill-posed because an infinite number of reflectance-shading combinations could produce the observed image color. Recent methods have successfully tackled this problem by incorporating priors on local shading [22, 33], enforcing sparse reflectance [10, 5], propagating user-specified constraints [7], leveraging reconstructed scene geometry [19, 1], or inferring constraints from multiple images with varying lighting [34, 12]. Like these latter approaches, we leverage multiple observations of the scene to constrain the decomposition.

In this paper, we describe a new method for intrinsic image decomposition, where the input is an image sequence of a static scene with fixed viewpoint under varying illumination. We show that the observation of temporal variations in local neighborhoods provides us with all the necessary information to relate the reflectance of multiple pixels, both at the local level and across pairs of distant pixels. This allows us to solve the decomposition problem and obtain intrinsic layers without the need of geometry, user interaction, an explicit lighting model, or assumptions on light color and reflectance sparsity.

We formulate the intrinsic decomposition as an optimization problem where our energies are derived directly from the Lambertian image formation model. Specifically, we make the following contributions:

- we derive a local energy that relates reflectance values within each neighborhood by analyzing the local color variations observed over time. This term is locally adaptive as it enforces smooth shading only on the relevant pixels, and automatically detects regions with sudden shading changes due to shadows or normal/depth discontinuities (Sec. 4);
- we infer pairwise constraints on the reflectance of distant pixels that are oriented consistently and have similar shading over time. We propose a robust approach for establishing reliable pairwise constraints without prior knowledge of the lighting or geometry (Sec. 5);
- we extend a synthetic benchmark for intrinsic image decomposition and conduct an extensive comparison with several recent methods (Sec. 6.2).

2. Related work

Intrinsic image decomposition is an ill-posed problem, and can only be solved by imposing constraints on the decomposition.

Constraints on the decomposition. Several assumptions on the shading or reflectance have been proposed. This paper does not attempt to exhaustively describe all the published methods, and we refer the reader to the work by Barron et al. [2] for a comprehensive review. Many methods share some common assumptions, e.g., smooth shad-

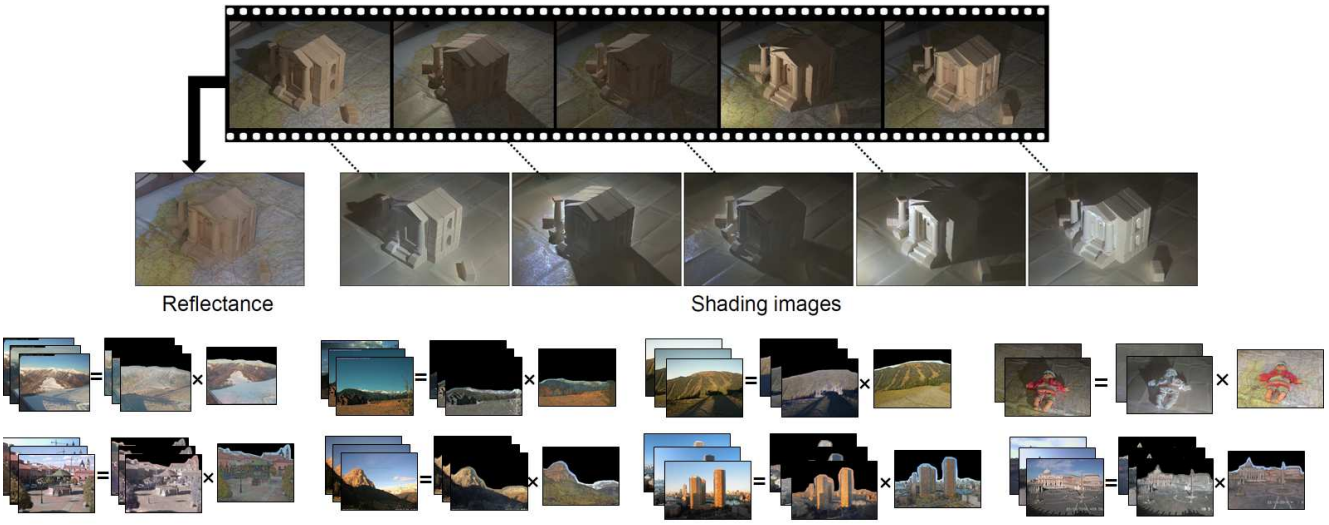


Figure 1. Top: given an input image sequence with fixed viewpoint and varying illumination, our method recovers the reflectance layer and a shading layer for each input frame. Despite specular effects, complex geometry and texture, our method properly recovers details in the reflectance layer while preserving smooth shading or hard shadows when required. Bottom: we ran our method on timelapses captured outdoors [21, 30], indoors [20], sequences from the MIT benchmark [11], and on the extended synthetic St. Basil dataset that we introduce in Section 6.2.

ing [22], sparse or piecewise-constant reflectance across the image [9], or grayscale shading [5]. They impose priors on the reflectance/shading values either by looking at local neighorhoods [7], distant relations between pixels [36], or global constraints to enforce gradient distributions [24] or sparsity [10] across the entire image. We show in this paper that we can decompose an input sequence without sophisticated priors on the reflectance or shading, by automatically learning where to enforce smooth shading and distant constraints based on multiple observations of the scene.

Additional input. Some methods leverage additional input to infer constraints. Bousseau et al. [7] and Bell et al. [5] allow individual or crowdsourced users to provide reflectance or shading annotations. Barron et al. [1], Laffont et al. [19, 20], and Hauagge et al. [13] leverage scene geometry, which is either captured with a depth sensor or reconstructed from multiple viewpoints. Bonneel et al. [6] decompose video sequences and let users interactively refine the results, while Ye et al. [35] propagate an existing decomposition to every frame of a video. Other methods decompose timelapses, i.e., sequences captured from a fixed viewpoint under varying lighting conditions. Such sequences are readily available for many outdoor scenes [14, 21, 30] or can be captured indoors with a fixed camera and a moving light source. Weiss et al. [34] use a prior derived from the image statistics of natural scenes, while Matshushita et al. [26] explicitly model temporal and spatial constraints. Hauagge et al. [12] propose a simplified physical model for modeling the local visibility, with a moving directional light source and constant ambient lighting. Sunkavalli et al. [31, 32] target outdoor scenes and model the lighting as a mixture of two light sources: directional sunlight and ambient skylight. Photometric stereo methods [4] use timelapses to solve the related problem of estimating per-pixel normals, but often ignore cast shadows which are an important component of shading. Our method also leverages the information contained in a time-lapse, but makes no assumption on the number or type of light sources; it can handle outdoor and indoor scenes, with an arbitrary number of light sources and shadows.

Evaluation. It is very difficult to gather ground truth reflectance and shading on real scenes, thus only few datasets are available to quantitatively evaluate the results of intrinsic image methods. The MIT dataset by Grosse et al. [11] is widely used, but consists of single objects that have been carefully captured in a lab environment and have been chosen to minimize interreflections. Real scenes are much more complex, as demonstrated in the crowdsourced database of Bell et al. [5]. Laffont et al. [20] proposed a synthetic dataset on a challenging scene created with physically-based rendering, for single-image or multi-view input. We extend this dataset to multiple lighting conditions for each view and use it for the evaluation of several recent methods.

3. Proposed approach

Our method takes as input a stack of T frames captured from a single viewpoint, under varying lighting. We denote by M the number of pixels in each frame, and by $\mathcal{I}_t$

the image at time $t \in \{1, \dots, T\}$. $I(p,t)^{(c)}$ represents the observed value of pixel p at frame t, in color channel c. Assuming that the scene is static and Lambertian, we write the intensity of each pixel p in image $\mathcal{I}_t$ as:

$$I(p,t)^{(c)} = R(p)^{(c)}\, S(p,t)^{(c)}. \tag{1}$$

In the following we ignore the superscript c and consider each channel independently, unless explicitly stated.

Our goal is to decompose the input image sequence into one reflectance image and T shading images (see Fig. 1). Our method proceeds as follows:

a) We first detect regions where shading is locally smooth, by focusing on each local neighborhood and analyzing its temporal variations across the sequence. In Sec. 4.1, we derive a local energy that relates reflectance values within each neighborhood.

b) We combine the local solutions from overlapping patches over the entire image to derive a locally adaptive term that enforces smooth shading in flat regions. In Sec. 4.2, we obtain an energy E_{local} that relates the reflectance of all pixels over the image.

c) We incorporate pairwise constraints to connect regions separated with depth/normal discontinuities. In Sec. 5, we derive an energy term E_{pair} that relates the reflectance of pairs of pixels that are consistently illuminated across the entire sequence.

We combine the energy terms E_{local} and E_{pair} into a total energy, which we minimize with respect to the reflectance. We can then derive the shading images by dividing each input frame by the estimated reflectance.

4. Local constraints from temporal color variations

In this section, we relate the reflectance values of pixels across the image by deriving a locally adaptive term enforcing smooth shading, based on the local color observations over time. We divide the input images into a set of overlapping patches Γ, which consists of all the square patches of size N pixels centered on every pixel.

In the following, we consider a single patch $\mathcal{W}_i \in \Gamma$. We build our smoothness energy on the observation that, under certain conditions, the shading over all pixels in $\mathcal{W}_i$ is constant. More specifically, we show in the supplementary material that the shading in Eq. 1 corresponds to the irradiance on a Lambertian surface. Assuming the surface is locally flat and that all points visible in $\mathcal{W}_i$ receive the same quantity of incident light in image $\mathcal{I}_t$, we can write for each pixel $p \in \mathcal{W}_i$:

$$I(p,t) = R(p)\, S(i,t). \tag{2}$$

When this equation holds, it allows us to relate the reflectance values of pixels within the patch, because they are all linked to a single shading value $S(i,t)$.

Eq. 2 assumes that the shading is locally constant within a small patch. Smooth shading is widely assumed in previous work [5, 10, 22, 29, 35, 36]; especially by Retinex-based methods to globally classify edges as reflectance or shading edges. In contrast, we enforce the smoothness assumption *locally*, only where it appears to hold based on the multiple observations, while relaxing it everywhere else.

In Sec. 4.1, we observe a single patch $\mathcal{W}_i$ under varying illumination and estimate in which images Eq. 2 holds. We define a local energy for this patch and mitigate the influence of violating frames in a robust manner, using Iteratively Reweighted Least Squares. This allows us to reduce the influence of these frames when relating local reflectance values. In Sec. 4.2, we then combine local energies from overlapping patches in order to relate the reflectance of all pixels across the entire image.

4.1. Local energy for planar surfaces

By following Eq. 2 and considering all the pixels in patch $\mathcal{W}_i \in \Gamma$, we define the local energy:

$$e_{\text{local}}(i,t) = \sum_{p \in \mathcal{W}_i} \left(R(p) - \frac{1}{S(i,t)} I(p,t) \right)^2. \tag{3}$$

For patch $\mathcal{W}_i$ and frame t, the unknowns are the shading $S(i,t)$, constant across the patch, and the reflectance $R(p)$ for each pixel in patch $\mathcal{W}_i$.

Unreliable images. To leverage the multiple observations of the patch $\mathcal{W}_i$ over time, a straightforward but inappropriate approach would be to simply sum this energy over all frames t. This local energy is based on the assumption from Eq. 2 that the shading is locally smooth, and if the assumption holds, the optimal lowest value of $e_{\text{local}}(i,t)$ at each frame t is 0. Nevertheless, it is important to note that this assumption is violated in two cases:

- when patch $\mathcal{W}_i$ contains a shadow boundary in image $\mathcal{I}_t$, Eq. 2 does not hold in the corresponding frame t;
- when patch $\mathcal{W}_i$ contains a normal or depth discontinuity, Eq. 2 does not hold for any of the input images.

Therefore, all the frames should not be considered in an equal way: the frames where the assumption is violated should "count" less for estimating the reflectance. These outliers will have a high residual and thus a large influence on the estimation. To reduce the influence of the outliers, we use a robust cost function $\rho()$. By considering the multiple observations over t in a robust way, we define:

$$e_{\text{local}}(i) = \sum_t \rho\left(e_{\text{local}}(i,t)\right). \tag{4}$$

The unknowns are *(1)* the reflectance $R(p)$ of every pixel in patch $\mathcal{W}_i$, that we stack into a $N \times 1$ column vector $\mathbf{R}_i$;

and *(2)* the shading values $S(i,t)$ of the patch $\mathcal{W}_i$ in every frame t, that we stack into a $T \times 1$ column vector $\mathbf{S}_i$. The choice of the robust cost function is discussed in the supplementary material. We use $\rho(x) = \sqrt{x}$ in all our results.

IRLS formulation. We minimize Eq. 4 by the Iteratively Reweighted Least Squares (IRLS) approach [27, 28, 18]. We will show that IRLS allows us to solve the equation robustly with a single linear system, but also eliminate unknowns so that the reflectance becomes the only unknown variable.

IRLS conducts robust model fitting by reweighting each data point in an iterative manner and by solving a weighted least squares problem at each iteration with the current weights. The weights control how much influence a data point has on the estimate of the model. Along the IRLS iterations, frames where Eq. 2 is violated are assigned smaller and smaller weights, and will in turn have a very small or negligible influence on the estimated model.

Each IRLS iteration consists of two steps: estimation of the model given the weights, and then update of the weights given the newly estimated model.

Model estimation: At each iteration k of IRLS, we estimate the model by solving:

$$(\mathbf{R}_i^{(k+1)}, \mathbf{S}_i^{(k+1)}) = \arg\min_{\mathbf{R}_i, \mathbf{S}_i} \sum_t w_{it}^{(k)} e_{\text{local}}(i,t) \qquad (5)$$

where $w_{it}^{(k)}$ are the weights obtained at the previous iteration, and $e_{\text{local}}(i,t)$ depends on $\mathbf{R}_i$ and $\mathbf{S}_i$. At the first iteration, we initialize $w_{it}^{(0)} = 1$ for all t.

Weights update: The residuals $e_{\text{local}}^{(k+1)}(i,t)$ computed with $\mathbf{R}_i^{(k+1)}$ and $\mathbf{S}_i^{(k+1)}$ obtained at Eq. 5 are now used to update the weights. Following [8], the weights are updated by $w_{it}^{(k+1)} = \rho'\left(e_{\text{local}}^{(k+1)}(i,t)\right)$ with $\rho'(x) = \partial\rho(x)/\partial x$.

Removing the shading unknown. The unknowns of Eq. 5 are both $\mathbf{R}_i$ and $\mathbf{S}_i$. We show in the supplementary material that within each patch $\mathcal{W}_i$, the shading can be expressed as a function of the reflectance by accumulating pixel values over the entire patch. In matrix form, we can write their relation:

$$\frac{1}{S(i,t)} = \mathbf{I}_{it}^{\intercal}\mathbf{R}_i / (\mathbf{I}_{it}^{\intercal}\mathbf{I}_{it}) \qquad (6)$$

where $\mathbf{I}_{it}$ is a $N \times 1$ column vector containing the stacked observed values of the N pixels in patch $\mathcal{W}_i$ in frame t. Plugging this into Eq. 3 allows us to express the local energy as a function of only the unknown $\mathbf{R}_i$. In matrix form, we can now rewrite Eq. 3 as:

$$e_{\text{local}}(i,t) = \|\mathbf{M}_{it}\mathbf{R}_i\|^2. \qquad (7)$$

where $\mathbf{M}_{it} = \mathbf{Id}_N - (\mathbf{I}_{it}^{\intercal}\mathbf{I}_{it})^{-1}\mathbf{I}_{it}\mathbf{I}_{it}^{\intercal}$ and with $\mathbf{Id}_N$ the identity matrix of size $N \times N$. Reinjecting this into Eq. 5, we obtain:

$$\mathbf{R}_i^{(k+1)} = \arg\min_{\mathbf{R}_i} \sum_t w_{it}^{(k)} \|\mathbf{M}_{it}\mathbf{R}_i\|^2. \qquad (8)$$

Minimizing the local energy. At each iteration of IRLS, the unknown in Eq. 8 is now only $\mathbf{R}_i$. Note that a trivial solution is $\mathbf{R}_i = 0$, and the solution is up to scale. Without lack of generality, we set the norm of $\mathbf{R}_i$ to 1 in the current subsection as our goal is to identify regions where we can impose shading smoothness. The optimization of Eq. 8 can be performed via SVD: the optimal solution $\mathbf{R}_i$ corresponds to the singular vector associated to the smallest singular value of the matrix $\mathbf{M}_i$ defined as

$$\mathbf{M}_i = \sum_t \sqrt{w_{it}^{(k)}}\mathbf{M}_{it}. \qquad (9)$$

Applying SVD is computationally cheap in our case because $\mathbf{M}_i$ is only of size $N \times N$ where $N = 9$ in all our results shown. We now have a robust estimate of the per-pixel reflectance (up to scale) within patch $\mathcal{W}_i$, based on the local color observations over the entire sequence.

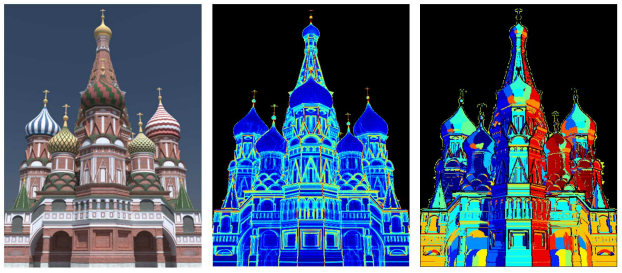
The output of this process is a matrix $\mathbf{M}_i$ that relates the reflectance values of all the pixels in patch $\mathcal{W}_i$. This matrix of size $N \times N$ is computed as a weighted sum (Eq. 9) where the weights estimated via IRLS are used to mitigate the influence of frames where the shading is not constant within the patch. This process enables us to *(1)* detect patches that lie on discontinuities (see Fig. 2b) and *(2)* ignore the incidental frames where the local shading is not smooth.

4.2. Locally adaptive energy

We have so far defined a local energy that relates the reflectance of pixels within one single patch, and have obtained weights that correspond to the reliability of our local estimates. We now relate the reflectance of all pixels across the image. We average the local energy of all overlapping patches in order to obtain the adaptive energy E_{local}:

$$E_{\text{local}} = \frac{1}{|\Gamma|}\sum_{\mathcal{W}_i \in \Gamma} e_{\text{local}}(i) = \frac{1}{|\Gamma|}\sum_{\mathcal{W}_i \in \Gamma} \|\mathbf{M}_i\mathbf{R}_i\|^2. \qquad (10)$$

This energy term relates the reflectance of all pixels in the image, since each pixel appears in multiple overlapping patches. Because we combine the local energies $e_{\text{local}}(i)$ obtained with our robust estimation (Sec. 4.1), this locally adaptive energy term selectively imposes a smooth shading in regions of the image where Eq. 2 holds, i.e. where the observed surfaces are diffuse and locally planar. The influence of occasional shadow boundaries or specularities in specific frames is mitigated since the estimated weight w_{it} is low in such cases. Regions with normal or depth discontinuities are detected (Fig. 2b) and their influence is mitigated.



(a) Input image **(b)** Residual of locally adaptive energy **(c)** Clusters from appearance profiles

Figure 2. Residual and clustering result from the input image. Left: input image. Middle: color-coded residual of our per-patch local energy $e_{\text{local}}(i)$; this energy term enforces smooth shading in regions of low residual (blue). Right: clusters from appearance profiles; each cluster contains pixels with similar shading in most images.

4.3. Discussion

We use multiple observations of the scene under different lighting in order to derive our locally adaptive energy term. By leveraging this extra information, we avoid making additional assumptions about the scene content (e.g., grayscale shading or priors on reflectance and shading) or tuning thresholds for classifying local gradients into reflectance or shading edges (e.g., Retinex variants).

Note that the smoothness energy used for intrinsic decomposition by Bousseau et al. [7] has a similar matrix form, but differs in two points. First, it aims to estimate the shading S in one single image, whereas our unknown is the reflectance R for an entire sequence. Second, it makes the assumption that reflectance values within each patch lie within a color plane. We make no such assumption; our local energy automatically detects patches and images where the shading is locally constant, and uses this information to constrain the reflectance values of the corresponding pixels.

5. Pairwise constraints from shading consistency

The locally adaptive energy proposed in Sec. 4.2 allows to constrain unknown reflectance values locally, in image regions with smooth surfaces. However, it cannot cross normal or depth discontinuities: in those areas, the residual from Eq. 3 is high and weights w_{it} are correspondingly low, thus preventing the local smoothness energy from constraining their relative reflectance values (see Fig. 2b). Using this local term E_{local} *alone* can produce weakly connected regions in terms of reflectance constraints (e.g., the dome areas in Fig. 2 are surrounded by boundaries with high residual) and can yield an intrinsic decomposition where the shading of separate surfaces is not consistent across the entire image.

In this section, we introduce *long-distance constraints* on pairs of pixels that can be distant in image space but whose intensities over time are related. These pairwise constraints allow us to relate the reflectance of regions that were only weakly connected with our locally adaptive smoothness term alone.

By definition of the intrinsic image model (Eq. 1), if the shading in frame t is the same for two pixels p and q, $S(p,t) = S(q,t)$. Then, from Eq. 1, we can write:

$$R(p)I(q,t) = R(q)I(p,t). \tag{11}$$

This means that if we could reliably find pairs of pixels that share a similar (yet unknown) shading in one or several frame(s) t, we would then be able to constrain the relative values of their reflectances. The key challenges are: *(1)* to identify pairs of pixels that have similar shading across most frames; *(2)* once good pairs of pixels have been identified, to mitigate the influence of occasional frames in which the two pixels accidentally have different shading (e.g., when one point is shadowed and not the other).

A similar approach was used by Laffont et al. [20], who constrained the reflectance ratio of two points based on the median of their radiance ratios observed over time. A key difference is that they use the normals reconstructed from multi-view input photographs in order to select pairs of points. In contrast, in Sec. 5.1, we select pairs without prior knowledge of the scene geometry, simply by analyzing their appearance profiles. This makes our method not only easier to apply but also considerably faster than [20]. We then constrain the relative reflectance of such pairs of consistent pixels, using the pairwise energy described in Sec. 5.2.

5.1. Selecting consistent pairs of pixels

We now select consistent pairs of pixels, which are likely to have similar shading in most images. To do this, we work on the appearance profile, that is the evolution of the intensity of each pixel p across frames: $[I(p,1), I(p,2), ..., I(p,T)]$. The intuition behind our approach is that, for a Lambertian scene, two points whose appearance profiles are proportional are likely to consistently have the same shading in several frames. A naive approach would consist in randomly sampling pairs of pixels and saving those with proportional appearance profiles; however, this would be computationally expensive as the vast majority of pixel pairs have different appearance profiles.

We propose an accelerated approach that is not only tractable but runs in just a few seconds. The main idea is to cluster pixels based on their appearance profiles: similar profiles correspond to points that have a similar evolution (up to scale), and thus similar shading in most images. We start by normalizing the appearance profiles, since shading is a scaled version of the pixels' profiles. We then perform

clustering with K-means on the normalized profiles; an example of the obtained clusters in Fig. 2c. A similar idea was described by Koppal et al. [17] to cluster surfaces based on the extrema of their appearance profiles; however, their goal is to find clusters of surfaces with similar *normals*, whereas we only aim to find pairs of pixels with similar *shading*.

Once the clusters have been found, we sample pairs of pixels within each cluster – since their appearance profiles are close to the cluster centroid, they will be close to each other as well. We sample a number of pairs per cluster, proportionally to the cluster size, so that we accumulate pairs of pixels over the entire image. The output of this process is a set Φ of pixel pairs, whose appearance profiles are similar over most frames.

5.2. Pairwise energy for consistently illuminated points

Assuming the two pixels of a pair $(p, q) \in \Phi$ have been properly chosen and they have the same (unknown) shading, a single frame (corresponding to a single t) is theoretically sufficient to infer a relative constraint between their reflectances according to Eq. 11. However, combining the information from multiple images with a robust cost function allows us to estimate how confident we are that both pixels share the same shading in multiple images. It also mitigates the influence of outlier frames where Eq. 11 incidentally does not hold, e.g. in an image where the two pixels are distant and separated by a shadow boundary.

Given pair $(p, q) \in \Phi$, we define the following energy:

$$e_{\text{pair}}(p, q) = \sum_t \left(R(p)I(q, t) - R(q)I(p, t)\right)^2 \qquad (12)$$

To handle outliers, we follow an approach similar to Sec. 4.1: we use a robust cost function $\rho()$ and apply IRLS to iteratively minimize the energy and update the weights. At each iteration k of IRLS, we optimize

$$(R(p)^{(k+1)}, R(q)^{(k+1)}) = \underset{R(p),R(q)}{\operatorname{argmin}} \sum_t w_{pqt}^{(k)} \left(R(p)I(q, t) - R(q)I(p, t)\right)^2 \qquad (13)$$

This can be rewritten in matrix form:

$$(\mathbf{R}_{pq}^{\text{pair}})^{(k+1)} = \underset{\mathbf{R}_{pq}^{\text{pair}}}{\operatorname{argmin}} \| \sqrt{\mathbf{W}_{pq}^{\text{pair}}} \mathbf{M}_{pq}^{\text{pair}} \mathbf{R}_{pq}^{\text{pair}} \|^2 \qquad (14)$$

where $\mathbf{W}_{pq}^{\text{pair}}$ is a diagonal $T \times T$ matrix containing the weights $w_{pqt}^{(k)}$ for all frames, $\mathbf{M}_{pq}^{\text{pair}} = [-\mathbf{I}_q, \mathbf{I}_p]$ a $T \times 2$ matrix that stacks the intensities of p and q, and $\mathbf{R}_{pq}^{\text{pair}} = [R(p); R(q)]$ a 2×1 vector that stacks the unknown reflectance of p and q.

The solution of Eq. 14 is obtained by SVD of $\sqrt{\mathbf{W}_{pq}^{\text{pair}}} \mathbf{M}_{pq}^{\text{pair}}$ which is extremely fast since this matrix is of size 2×2 for each pair.

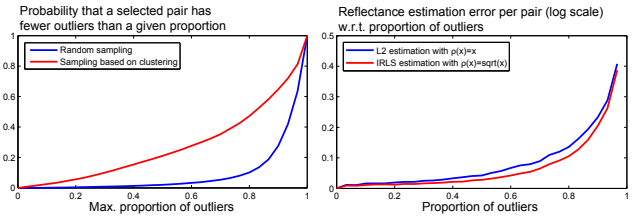


Figure 3. The approach described in Sec. 5.1 aims to select pairs of pixels with similar shading in most images; the cumulative distribution function (left) shows that this method yields pairs of pixels with significantly fewer outliers than random sampling. Given pairs of pixels with some outlier frames (right), the IRLS approach described in Sec. 5.2 yields more accurate pairwise reflectance ratios than by directly minimizing Eq. 13; we define error as $\left| log(\frac{R(p)}{R(q)}) - log(\frac{\widetilde{R}(p)}{\widetilde{R}(q)}) \right|$ with $\widetilde{R}$ the ground-truth reflectance.

Finally, we average the pairwise energy over all the pairs selected in Sec. 5.1, yielding the energy E_{pair} that constrains the reflectance of a sparse set of pixels over the entire image:

$$E_{\text{pair}} = \frac{1}{|\Phi|} \sum_{(p,q)\in\Phi} e_{\text{pair}}(p, q) \qquad (15)$$

6. Implementation and results

We define a global energy by summing E_{local} and E_{pair} weighted by a scalar parameter γ_{pair}. We aim to minimize this energy with respect to the reflectance:

$$\underset{\mathbf{R}}{\operatorname{argmin}}\, E_{\text{local}} + \gamma_{\text{pair}} E_{\text{pair}}. \qquad (16)$$

This translates into a large sparse linear system of the form $Ax = 0$. Such a system could theoretically be solved via SVD, but this would be prohibitive given the size of the images (matrix A has as many rows and columns as the total number of pixels in the image). Instead, we add a regularization term $\gamma_{\text{reg}} E_{\text{reg}}$, where $E_{\text{reg}} = \sum_p \sum_t \left(R(p) - \frac{3I(p,t)}{\sum_c I(p,t)^{(c)}} \right)^2$. This commonly used term favors reflectance values close to the chromaticity of the input image. As this is not always reliable for real scene images, we use a tiny weight γ_{reg} so it does not influence the solution.

We solve the global sparse linear system and obtain the per-pixel reflectance. We repeat this operation separately in each color channel. Finally, we obtain the shading in each frame by dividing each input by the common reflectance.

Pairwise constraints sampling and estimation. The method described in Sec. 5.1 aims to select consistent pairs of pixels which are likely to have similar shading in most images. This selection process is very challenging since shading is unknown beforehand. We show in Fig. 3 (left)

that our approach based on clustering appearance profiles selects pairs of pixels that are more likely to have similar shading in some images – in other words, they have fewer *outlier* frames which violate Eq. 11. For the visualization in Fig. 3, we consider frame t to be an outlier for pair (p, q) if the ground-truth shading values $\widetilde{S}(p,t)$ and $\widetilde{S}(q,t)$ differ by 5% or more; we access ground truth values by using the synthetic scene shown in Fig. 2.

Once a reasonable set of pairs has been identified, the robust approach described in Sec. 5.2 is used to mitigate the influence of outlier frames and estimate pairwise reflectance ratios. Fig. 3 (right) compares the mean error in reflectance estimation achieved with two robust function $\rho()$, for pairs with varying proportions of outlier frames. We use the L^1 cost (combined with IRLS) to generate our results, as it is more robust to outliers than the standard L^2 cost. Note that unreliable pairs with a higher proportion of outliers also affect less the optimization of Eq. 16, as they will be associated with smaller weights.

We now present results on some popular synthetic and real datasets. We invite the reader to refer to our project webpage[1] for our extended St. Basil dataset (Sec. 6.2), additional results, figures, derivations, and details including values for our parameters.

6.1. Captured scenes: MIT benchmark

We applied our approach on the popular MIT intrinsic image dataset [11], for which the ground truth shading is approximately known (ignoring interrefections). We quantitatively evaluate several methods by comparing their results to the ground truth, using the Local Mean Squared Error (LMSE) described by Grosse et al. [11]. The quantitative comparison is available in Fig. 4. Our method provides very satisfying results and favorably compares to state-of-the-art methods. Examples of mean runtimes of existing methods as reported in the literature are: [10]: >600s, [29]: >300s, [33]: >200s, [2]: >200s, [16]: 40s, [36]: 3s, [24]:1-3s, Retinex [11]: 1s. On our PC equipped with a Core i5 2.6 Ghz CPU, [12] runs in 257s and ours in 94s on average.

On this dataset, only the method by Hauagge et al. [12] achieves better LMSE results. Note however, that the MIT dataset only contains isolated objects captured in laboratory conditions, with grayscale shading and no interreflections. Although we evaluate our results on this dataset, our method targets more complex scenes and does not make such assumptions. In addition, many objects from the dataset are biased towards methods with a sparse reflectance prior, since they are mostly white with a few hand-drawn colored strokes.

[1] https://graphics.ethz.ch/~plaffont/research/intrinsicTimelapse/

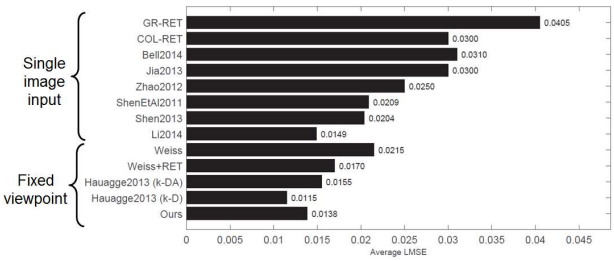


Figure 4. Comparison of Local Mean Squared Error (LMSE) on the MIT intrinsic dataset.

6.2. Synthetic dataset: Extended St. Basil

Several papers acknowledge the need for datasets with ground truth reflectance and shading in order to evaluate intrinsic image algorithms [10, 23, 15]. We propose the first synthetic dataset that depicts a scene with complex geometry, under multiple physically-based lighting conditions for each viewpoint, with ground truth reflectance and shading images. As shown in our extensive comparison, this dataset is challenging even for recent and state-of-the-art methods that perform well on the MIT benchmark (Sec. 6.1).

We extend the St. Basil dataset released by Laffont et al. [20], which depicts a complex 3D scene with a physically-based outdoor lighting model, under varying lighting conditions and viewpoints. From each of the three viewpoints corresponding to the evaluation images, we generate 30 input images with varying lighting and fixed camera position. This allows us to evaluate methods based on a timelapse input. We quantitatively evaluate 16 different methods using the Local Mean Squared Error (LMSE) and Global Mean Squared Error (GMSE) [11]. All the result images, as well as the evaluation scores for different techniques, are available on our project webpage.

Quantitatively, our method scores second on this benchmark, both in LMSE and GMSE (Fig. 5). It is overperformed only by Laffont et al. [20], which use a different input compared to our method: since they use images from multiple viewpoints, they have information about the scene geometry and use it to compensate for ambient occlusion. The method by Weiss [34] yields an LMSE similar to ours, but its GMSE is the highest of all the tested approaches; this is likely due to the reintegration step and yields a result is visually very different from ground truth. In contrast, our long-distance pairwise constraints enforce consistent shading across multiple surfaces in the scene.

Our method yields visually pleasing results. Compared to ground truth, most of the remaining artifacts are in the lower part of the building, under the arches. This is because most of the images in this dataset were captured during daytime and illuminated from the top, thus most of the shading comes from indirect lighting in those areas. These

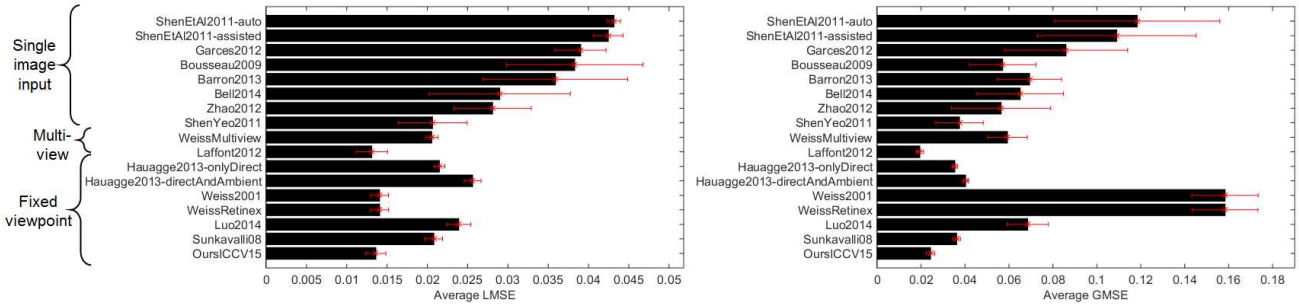


Figure 5. Quantitative comparison of LMSE (left) and GMSE (right) for several methods on the extended synthetic St. Basil dataset.

artifacts are shared with the other timelapse-based methods, including the one from Sunkavalli et al. [32] which explicitly models outdoor lighting.

6.3. Captured scenes: Doll, Temple, and timelapses

We provide results obtained by our method on two indoor datasets Temple and Doll from [20] shown in Fig. 1, both containing 7 images. The input data have been captured by a camera on a tripod and with a moving light source. Note that these datasets are much more complex since they exhibit intricate textures, strong lighting, specular surfaces and other non-Lambertian effects (e.g. on the doll's clothes). Most single-image based methods perform poorly in these cases, as the current most-advanced methods cannot disambiguate reflectance from shading on the textured tablecloth. Nevertheless, the shading layer we recover is both uncorrupted and smooth, and at the same time our method is still able to capture fine shading details such as the foldings of the tablecloth.

Fig. 1 shows results of our decomposition on seven timelapse sequences captured outdoors. More results and comparisons are available on our project webpage.

6.4. Discussion and limitations

Our model makes no assumption on the number or type of light sources; it can handle outdoor and indoor scenes (see Fig. 1), shadows (including from nearby light sources, as in our indoor scenes), interreflections (see the base of the domes in St. Basil).

Our model assumes static Lambertian scenes. Since the reflectance layer is fixed for each sequence, non-Lambertian effects (e.g., specular highlights) and moving objects are assigned to the shading layer. Although our method is not designed to deal with a high amount of specularities, it can cope with such view-dependent effect to some extent as shown in our results (e.g. shiny map in indoor scenes of Fig. 1).

When all surfaces are orthogonal, no pairwise constraints can be established between pixels across different surfaces. This can occur on scenes composed of only one isolated and matted box, e.g., the Box scene in the MIT dataset. In practice however, many surfaces tend to have similar (yet not identical normals) in a real scene, thus allowing all surfaces to be connected to each other.

7. Conclusion

We presented a method to compute the intrinsic image decomposition from an image sequence acquired by a static camera. Our approach runs automatically and returns the shading layer of each of the input images, as well as the reflectance. We defined a locally adaptive energy from the observations of each local neighborhood over time, and then enforced distant pairwise constraints across surfaces shaded consistently. We demonstrated the validity of our approach with extensive comparisons on both synthetic and real datasets, and showed that our approach favorably compares with state-of-the-art methods.

Acknowledgments

This research is supported by the BeingThere Centre, a collaboration between Nanyang Technological University (NTU) Singapore, Eidgenössische Technische Hochschule (ETH) Zurich, and University of North Carolina (UNC) at Chapel Hill. The BeingThere Centre is supported by the Singapore National Research Foundation (NRF) under its International Research Centre @ Singapore Funding Initiative and administered by the Interactive Digital Media Programme Office (IDMPO). The authors thank Kalyan Sunkavalli, Jean-François Lalonde, YiChang Shih, and the authors of concurrent methods for sharing their results.

References

[1] J. T. Barron and J. Malik. Intrinsic scene properties from a single RGB-D image. In *CVPR*, 2013. 1, 2

[2] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *TPAMI*, 2015. 1, 7

[3] H. Barrow and J. Tenenbaum. Recovering intrinsic scene characteristics from images. *Computer Vision Systems*, 1978. 1

[4] R. Basri, D. W. Jacobs, and I. Kemelmacher. Photometric stereo with general, unknown lighting. *IJCV*, 2007. 2

[5] S. Bell, K. Bala, and N. Snavely. Intrinsic images in the wild. *TOG (SIGGRAPH)*, 2014. 1, 2, 3

[6] N. Bonneel, K. Sunkavalli, J. Tompkin, D. Sun, S. Paris, and H. Pfister. Interactive intrinsic video editing. *TOG (SIGGRAPH Asia)*, 2014. 2

[7] A. Bousseau, S. Paris, and F. Durand. User-assisted intrinsic images. *TOG (SIGGRAPH)*, 2009. 1, 2, 5

[8] O. Enqvist, F. Kahl, and R. Hartley. Robust optimization techniques in computer vision. *ECCV tutorial*, 2014. 4

[9] E. Garces, A. Munoz, J. Lopez-Moreno, and D. Gutierrez. Intrinsic images by clustering. *Computer Graphics Forum (proc. EGSR)*, 2012. 2

[10] P. Gehler, C. Rother, M. Kiefel, Z. Lumin, and B. Schoelkopf. Recovering intrinsic images with a global sparsity prior on reflectance. In *NIPS*, 2011. 1, 2, 3, 7

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

[12] D. Hauagge, S. Wehrwein, K. Bala, and N. Snavely. Photometric ambient occlusion. In *CVPR*, 2013. 1, 2, 7

[13] D. Hauagge, S. Wehrwein, P. Upchurch, K. Bala, and N. Snavely. Reasoning about photo collections using models of outdoor illumination. In *BMVC*, 2014. 2

[14] N. Jacobs, N. Roman, and R. Pless. Consistent temporal variations in many outdoor scenes. In *CVPR*, 2007. 2

[15] J. Jeon, S. Cho, X. Tong, and S. Lee. Intrinsic image decomposition using structure-texture separation and surface normals. In *ECCV*, 2014. 7

[16] K. Jia, X. Tang, and X. Wang. Image transformation based on learning dictionaries across image spaces. *TPAMI*, 2013. 7

[17] S. J. Koppal and S. G. Narasimhan. Appearance derivatives for isonormal clustering of scenes. *TPAMI*, 2009. 6

[18] V. Kwatra, I. Essa, A. Bobick, and N. Kwatra. Texture optimization for example-based synthesis. *TOG (SIGGRAPH)*, 2005. 4

[19] P.-Y. Laffont, A. Bousseau, and G. Drettakis. Rich intrinsic image decomposition of outdoor scenes from multiple views. *TVCG*, 2013. 1, 2

[20] P.-Y. Laffont, A. Bousseau, S. Paris, F. Durand, and G. Drettakis. Coherent intrinsic images from photo collections. *TOG (SIGGRAPH Asia)*, 2012. 1, 2, 5, 7, 8

[21] J.-F. Lalonde, A. A. Efros, and S. G. Narasimhan. Webcam clip art: Appearance and illuminant transfer from time-lapse sequences. *TOG (SIGGRAPH Asia)*, 2009. 2

[22] E. H. Land and J. J. Mccann. Lightness and retinex theory. *Journal of the Optical society of America*, 1971. 1, 2, 3

[23] K. J. Lee, Q. Zhao, X. Tong, M. Gong, S. Izadi, S. U. Lee, P. Tan, and S. Lin. Estimation of intrinsic image sequences from image+depth video. In *ECCV*, 2012. 7

[24] Y. Li and M. S. Brown. Single image layer separation using relative smoothness. In *CVPR*, 2014. 2, 7

[25] X. Liu, L. Wan, Y. Qu, T.-T. Wong, S. Lin, C.-S. Leung, and P.-A. Heng. Intrinsic colorization. *TOG (SIGGRAPH Asia)*, 2008. 1

[26] Y. Matsushita, S. Lin, S. B. Kang, and H.-Y. Shum. Estimating intrinsic images from image sequences with biased illumination. In *ECCV*, 2004. 2

[27] F. Pighin and J. P. Lewis. Practical least-squares for computer graphics. In *ACM SIGGRAPH Courses*, 2007. 4

[28] S.Bouaziz, A.Tagliasacchi, and M.Pauly. Dynamic 2d/3d registration. *EUROGRAPHICS Tutorial*, 2014. 4

[29] L. Shen, C. Yeo, and B.-S. Hua. Intrinsic image decomposition using a sparse representation of reflectance. *TPAMI*, 2013. 3, 7

[30] Y. Shih, S. Paris, F. Durand, and W. T. Freeman. Data-driven hallucination of different times of day from a single outdoor photo. *TOG (SIGGRAPH Asia)*, 2013. 2

[31] K. Sunkavalli, W. Matusik, H. Pfister, and S. Rusinkiewicz. Factored time-lapse video. *TOG (SIGGRAPH)*, 2007. 2

[32] K. Sunkavalli, F. Romeiro, W. Matusik, T. Zickler, and H. Pfister. What do color changes reveal about an outdoor scene? In *CVPR*, 2008. 2, 8

[33] M. F. Tappen, W. T. Freeman, and E. H. Adelson. Recovering intrinsic images from a single image. *TPAMI*, 2005. 1, 7

[34] Y. Weiss. Deriving intrinsic images from image sequences. In *ICCV*, 2001. 1, 2, 7

[35] G. Ye, E. Garces, Y. liu, Q. Dai, and D. Gutierrez. Intrinsic Video and Applications. *TOG (SIGGRAPH)*, 2014. 2, 3

[36] Q. Zhao, P. Tan, Q. Dai, L. Shen, E. Wu, and S. Lin. A closed-form solution to retinex with nonlocal texture constraints. *TPAMI*, 2012. 2, 3, 7