Intrinsic Video

Naejin Kong, Peter V. Gehler, and Michael J. Black

Max Planck Institute for Intelligent Systems, Tübingen, Germany
{naejin.kong,peter.gehler,black}@tuebingen.mpg.de

Abstract. Intrinsic images such as albedo and shading are valuable for later stages of visual processing. Previous methods for extracting albedo and shading use either single images or images together with depth data. Instead, we define *intrinsic video* estimation as the problem of extracting temporally coherent albedo and shading from video alone. Our approach exploits the assumption that albedo is constant over time while shading changes slowly. Optical flow aids in the accurate estimation of intrinsic video by providing temporal continuity as well as putative surface boundaries. Additionally, we find that the estimated albedo sequence can be used to improve optical flow accuracy in sequences with changing illumination. The approach makes only weak assumptions about the scene and we show that it substantially outperforms existing single-frame intrinsic image methods. We evaluate this quantitatively on synthetic sequences as well on challenging natural sequences with complex geometry, motion, and illumination.

Keywords: intrinsic images, video, temporal coherence, optical flow.

1 Introduction

Albedo and shading are fundamental view-centric features of the visual world that are closely related to physical properties of surfaces and light [5]. Many computer vision algorithms work directly with pixel values, which are a combination of albedo and shading. These same algorithms, whether image/video segmentation, flow estimation, object detection, or 3D shape reconstruction, may work significantly better if applied to more physical quantities. Consequently we seek to decompose images into their intrinsic albedo and shading components. This problem has been studied since the 1970's but previous work focuses on individual images. Here we extend these methods to video sequences and show that, by exploiting temporal constraints on albedo and shading, our method outperforms single-frame methods.

While today "intrinsic images" are typically taken to mean shading and albedo, the original meaning of Barrow and Tenenbaum [5] includes additional "images" related to object shape, such as surface normals, depth, and occluding contours. By using sequences of images, rather than static images, we extract a richer set of intrinsic images that include: albedo, shading, optical flow, occlusion regions, and motion boundaries. Our formulation provides an integrated framework for modeling video sequences in terms of such intrinsic images.

D. Fleet et al. (Eds.): ECCV 2014, Part II, LNCS 8690, pp. 360–375, 2014.

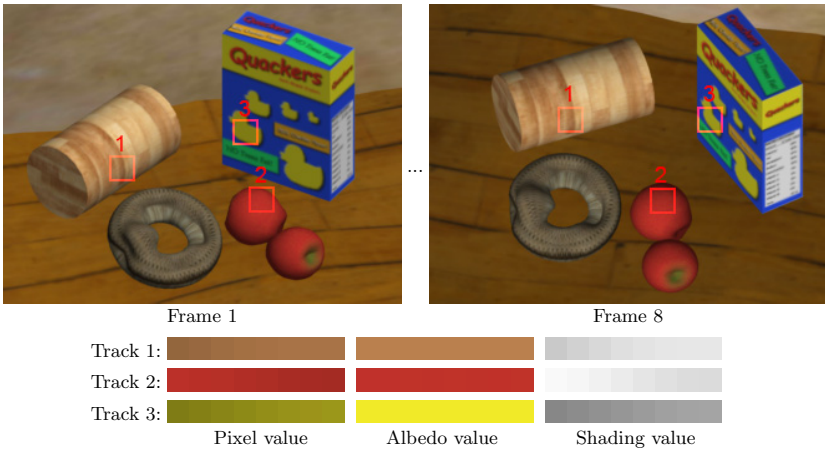

Fig. 1. Intrinsic Video. Top: two frames from a synthetic sequence with camera motion and illumination change. Bottom: pixel, albedo, and shading values for marked locations in 8 consecutive frames. Pixel values and shading change over time, while albedo is constant.

For a Lambertian surface, albedo and shading are mixed and encoded in observed pixel values according to

$$i_t(\mathbf{x}) = a_t(\mathbf{x}) \times s_t(\mathbf{x}), \tag{1}$$

where i is the known image, a and s are unknown albedo and shading variables, respectively, $\mathbf{x}$ is pixel location, and t is time. Since there are two unknowns and one observation, the recovery (factorization) of albedo and shading is ill-posed at a single pixel. To recover the intrinsic images from a single image, there have been several proposals for priors on albedo and shading that show promising results [4,12]. Previous work, however, has typically not considered videos of general scenes, non-rigid motions, and changing illumination. Our experiments with single-frame methods show that they do not produce temporally coherent results when applied independently to video frames.

If we know the optical flow of the scene, then we actually have additional constraints on the albedo and shading. If a surface is changing orientation with respect to the illumination, then the image values change, but the albedo does not. Thus, correspondence in time can provide additional constraints that make solving for the albedo well posed. We define *intrinsic video* as the factorization of video into sequences of albedo, shading, motion, occlusion, and motion boundaries.

Consider the synthetic image sequence in Fig. 1, containing camera motion and changing illumination. Given optical flow, the change in albedo and shading over time can be physically motivated. First, albedo is a unique value for each material that determines surface color, and its value is constant as long as the material stays unchanged. Second, shading is generated from physical

interaction between surface geometry and incoming light. It is reasonable to assume that a camera, or objects in the scene, move or transform smoothly while the lighting condition changes only a little within a short time interval. Then, each scene point will exhibit constant albedo but smoothly (and slowly) varying shading over the video sequence. We use these insights to formulate new priors for intrinsic video estimation. Of course, these assumptions are sometimes violated (e.g., by cast shadows) and we address this below using a robust statistical formulation.

Note also that if pixel values change over time, they violate the assumption of brightness constancy often used in the computation of optical flow [15]. This can cause optical flow algorithms to fail unless they are made robust to such changes [7]. Since albedo is constant, however, we show that it can be used to more accurately estimate optical flow (cf. [10]). We thus suggest that intrinsic video provides a framework for combining optical flow and intrinsic image estimation in a mutually beneficial way.

Specifically, our intrinsic video method uses optical flow to establish a temporal constancy term for albedo and a temporal smoothness term for shading. Our spatial priors on albedo are similar to those suggested in [4]; these encourage the estimated albedo to be sparse and uniformly smooth. We develop a non-local spatial prior on shading that encourages spatial smoothness of the estimated shading based on a median of local and non-local pixel neighbors. Optical flow also provides us with information about the structure of the scene that we can use to improve intrinsic image estimation from video. In estimating shading, we use geometric information available in the flow, such as motion boundaries and occlusion, to enhance the quality of the estimated shading images. The full solution uses the Classic+NL flow algorithm [26] as a foundation and extends it for intrinsic video estimation.

We show results on synthetic and real sequences with complex motions and illumination change. Previous datasets for static intrinsic image evaluation are not appropriate so we develop a new synthetic dataset that we make publicly available. Both quantitatively on synthetic sequences, and qualitatively on real sequences, we substantially outperform single-frame methods on the estimation of albedo and shading.

2 Previous Work

A classic approach to constrain albedo and shading estimation from a single image is based on the Retinex theory [19], which says that albedo edges tend to be stronger than shading edges. Its usefulness was first proved in [16]. The performance of Retinex-based algorithms depends on correct labeling of albedo and shading edges. Learning-based approaches automatically determine this labeling [6,28], or directly predict shading edges [27]. Grosse et al. [13] conduct a quantitative analysis using their ground truth dataset and find that Retinex-based approaches perform well (in 2009). The dataset, however, contains static images of single segmented objects.

Recent Retinex-based algorithms add more constraints on albedo or shading. Bousseau et al. [8] require a small number of user-labeled albedo and shading pixels. In [25] and [24], non-local texture cues or a local continuity assumption on albedo are used, respectively. Gehler et al. [12] develop a probabilistic model and add a new global sparsity prior on albedo that models natural image statistics.

Intrinsic images are related to the physics of image formation. Barron and Malik [2,3] exploit the physics by explicitly modeling the shape and lighting that generate shading. Their shape priors, however, cannot model a whole scene involving multiple objects and surface discontinuities. This limits their scope to single objects, pre-segmented from the single image. Their extension in [4] jointly estimates several depths and lights given depth constraints and estimated depths (e.g. from a range scanner). In [11], shading estimation is constrained by relying on surface points reconstructed from given depth. Current depth sensors are still noisy, however, and for archival images and videos, no explicit depth information exists. Scene structure, however, is *implicit* in an RGB video sequence and we show that the optical flow already contains enough approximate geometric information to estimate albedo and shading in scenes with multiple objects. In particular, optical flow allows us to extract occlusion regions and putative surface boundaries. We exploit these in estimating piecewise-smooth shading.

Lee et al. [20] extract intrinsic images from an RGB-D video. Their temporal constraints are mainly built upon pixel correspondences obtained from 3D coordinates reconstructed with depth. Laffont et al. [17,18] used a collection of photographs that capture the same scene from different views under varying illumination. These methods also need pixel correspondences including their normals across the photographs, which are obtained by applying multi-view stereo; this assumes a rigid scene. Our approach uses optical flow instead, thus making no strong assumption about the scene structure. By not assuming a rigid scene, our intrinsic video method can deal with non-rigid and independently moving objects. In addition, optical flow is useful for imposing image-based temporal constraints, since it provides dense pixel correspondences at the image level. Weiss [29] and Hauagge et al. [14] estimate a single albedo image from a series of images of the same scene captured under significantly varying lighting conditions. These methods do not work if anything in the scene moves or if there is camera motion; each pixel in the image series should represent a single point and contain as much light variation as possible.

Like us, Ye et al. [30] extract coherent intrinsic images from an RGB video. They use optical flow to propagate an initial albedo decomposition of the first frame over the video sequence. In contrast, our model comes from physical properties of visible surfaces under motion and illumination variation. We optimize a full objective function containing both shading and albedo that integrates priors on each, including the spatial albedo priors from [4], new temporal priors on albedo and shading in Section 3.2, and new spatial shading priors that approximate object boundaries in Section 3.3. Using optical flow, the approach integrates information throughout all frames in the video and extracts additional intrinsic images related to occlusions and motion boundaries.

3 Formulation

Given a sequence of images, $\{I_t\}$, we extract the *intrinsic video* sequence, $\{A_t, S_t\}$, of albedo and shading images at each time instant t. Unless otherwise specified, and without loss of generality, all images are in the log domain. We recover the intrinsic video sequence by minimizing this objective function

$$\underset{\{A_t,S_t\}}{\operatorname{argmin}} E(\{A_t, S_t\} \mid \{I_t\}, \{\mathbf{u}_t\}) =$$
$$\sum_t f_D(A_t, S_t|I_t) + f_{T_A}(A_{t+1}, A_t|\mathbf{u}_t) + f_{T_S}(S_{t+1}, S_t|\mathbf{u}_t) + f_A(A_t) + f_S(S_t), \quad (2)$$

where $\mathbf{u}_t$ is the optical flow between input images I_t and I_{t+1}. This flow is pre-computed from the input video using Classic+NL [26]. We assume that the estimated flow establishes reasonably accurate pixel correspondences robust to some illumination variation in the video.

The data term, $f_D(\cdot)$, enforces similarity between the input and reconstructed images (Section 3.1). The temporal coherence terms, $f_{T_A}(\cdot)$ and $f_{T_S}(\cdot)$, are pixel-wise temporal constraints on albedo and shading, respectively (Section 3.2); the formulation of these is one of our key novelties and goes beyond previous work. The spatial terms, $f_A(\cdot)$ and $f_S(\cdot)$, are priors, based on the statistics of albedo and shading, that constrain the solution (Section 3.3 and 3.4). Our spatial shading prior exploits optical flow in a novel way. Note that we assume the illuminant is white and thus shading is a grayscale image that has the same effect in each RGB channel. While the images are RGB, we often drop the index over RGB for clarity. Each term is described in detail below.

3.1 Image Similarity

The Lambertian equation (1), in the log domain, defines the data term. It measures similarity between each input log-image and the reconstructed log-image:

$$f_D(A_t, S_t) = \lambda_D \sum_{c\in\{\mathrm{R,G,B}\}} \sum_{\mathbf{x}} \rho_D \left(w_t^{\mathrm{lum}}(\mathbf{x})(I_t(\mathbf{x}, c) - A_t(\mathbf{x}, c) - S_t(\mathbf{x})) \right), \quad (3)$$

where $\mathbf{x}$ is pixel location, c is color channel, $w^{\mathrm{lum}}(\mathbf{x}) = \mathrm{lum}(i(\mathbf{x})) + \varepsilon$, and $\mathrm{lum}(i)$ takes the luminance from the input intensity image i and $\varepsilon = 0.001$. This weight has been proven to be useful in [11] to prevent disproportionally strong contributions of dark pixels. The function $\rho_D(\cdot)$ penalizes differences between the observed and predicted log image. To deal with violations of our assumptions, we use a robust Charbonnier function $\rho_{\mathrm{Charb}}(x) = \sqrt{x^2 + \epsilon^2}$ (a differentiable variant of the L1 penalty [9]; $\epsilon = 0.001$). The weight $\lambda_D = 10$ in all experiments.

3.2 Temporal Constraints

Inspired by Barrow and Tenenbaum [5], we formulate the intrinsic video problem to exploit physical properties of albedo and shading on the visible surfaces

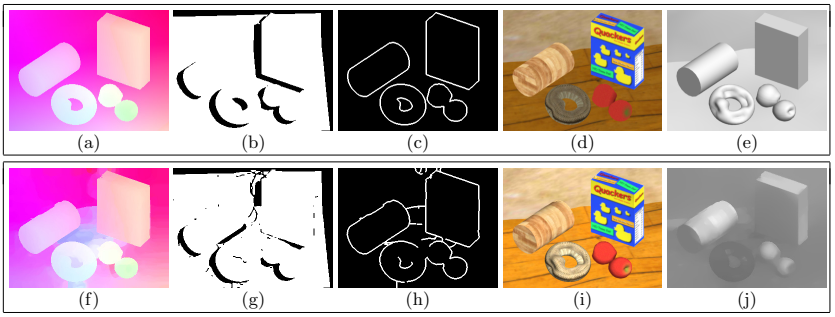

Fig. 2. Five types of intrinsic images. The occlusion map $w^{\text{occ}}_{\mathbf{u}_t}$ (Section 3.2) and the boundary map $w^{\text{bnd}}_{\mathbf{u}_t}$ (Section 3.3) are detected from optical flow. (a) Ground truth flow. (b) $w^{\text{occ}}_{\mathbf{u}_t}$ detected from (a). (c) $w^{\text{bnd}}_{\mathbf{u}_t}$ detected from (a). (d) Ground truth albedo. (e) Ground truth shading. (f) Estimated flow. (g) $w^{\text{occ}}_{\mathbf{u}_t}$ detected from (f). (h) $w^{\text{bnd}}_{\mathbf{u}_t}$ detected from (f). (i) Albedo estimated using (f)-(h). (j) Shading estimated using (f)-(h).

under motion and illumination variation. Specifically, we assume that albedo is typically constant over time while shading information changes slowly. Exploiting these assumptions requires that we know the correspondence of pixels over time. This is given by the optical flow, $\{\mathbf{u}_t\}$, over the sequence.

Temporal albedo constancy is defined as

$$f_{T_A}(A_t, A_{t+1}) = \lambda_{T_A} \sum_{c\in\{\text{R,G,B}\}} \sum_{\mathbf{x}} w^{\text{occ}}_{\mathbf{u}_t}(\mathbf{x}) \cdot \rho_{T_A}\Big(A_{t+1}(\mathbf{x}+\mathbf{u}_t(\mathbf{x}), c) - A_t(\mathbf{x}, c)\Big), \quad (4)$$

where $A_{t+1}(\mathbf{x}+\mathbf{u}_t(\mathbf{x}))$ represents the albedo warped by the optical flow and where $w^{\text{occ}}_{\mathbf{u}_t}$ is a weight map computed from the optical flow that is 0 if the pixel is occluded and 1 otherwise. This weight map is a type of intrinsic image that disables temporal coherence of albedo at occlusion boundaries; see Fig. 2 (g).

The choice of penalty function, ρ_{T_A}, is critical. Although a pixel $A_t(\mathbf{x})$ and a warped pixel $A_{t+1}(\mathbf{x}+\mathbf{u}_t(\mathbf{x}))$ should have the same values theoretically, they are in practice similar but not strictly equal due to aliasing and finite image sampling. Also any errors in the optical flow could lead to errors in albedo because the pixels do not correspond to the same physical location in the scene. Consequently we adopt the smooth but robust Tukey function:

$$\rho_{T_A}(x) = \rho_{\text{Tukey}}(x) = \begin{cases} \frac{1}{3} & \text{if } x < -\alpha \text{ or } x > \alpha \\ \frac{x^2}{\alpha^2} - \frac{x^4}{\alpha^4} + \frac{x^6}{3\alpha^6} & \text{otherwise,} \end{cases} \quad (5)$$

where $\alpha = 5$. This function is robust to various outliers caused by sampling, brightness variation, complex motion, occlusion and noise; it is also differentiable.

Optical flow is by nature undetermined in an image region occluded in the next image. The occlusion map $w^{\text{occ}}_{\mathbf{u}_t}$ is useful to prevent minor image artifacts where the flow is not defined. We detect occlusions by using the difference of input

images and the divergence of the optical flow. We threshold this and exclude pixels moving outside the image boundaries:

$$w^{\text{occ}}_{\mathbf{u}_t}(\mathbf{x}) = \begin{cases} 1, \text{ if } o(\mathbf{x}) \geq 0.5 \text{ and } \mathbf{x}+\mathbf{u}(\mathbf{x}) \text{ stays inside the image} \\ 0, \qquad\qquad\qquad\qquad \text{otherwise,} \end{cases} \tag{6}$$

where

$$o(\mathbf{x}) = \exp\left(-\frac{(i^{\text{Lab}}_t(\mathbf{x}) - i^{\text{Lab}}_{t+1}(\mathbf{x}+\mathbf{u}))^2}{2\sigma_e^2} - \frac{d^2_{\mathbf{u}_t}(\mathbf{x})}{2\sigma_d^2} \right), \tag{7}$$

$\sigma_d = 0.3$, $\sigma_e = 20$, and i^{Lab} is the input image in the Lab space. $d_{\mathbf{u}_t}$ is one-sided divergence computed from the flow $\mathbf{u}_t$. A similar detection heuristic is used in [23,26] for disabling the spatial regularization of optical flow and works well in our experiments.

Temporal shading similarity is defined as

$$f_{T_S}(S_t, S_{t+1}) = \lambda_{T_S} \sum_{\mathbf{x}} w^{\text{occ}}_{\mathbf{u}_t}(\mathbf{x}) \cdot \rho_{T_S}\left(S_{t+1}(\mathbf{x}+\mathbf{u}_t(\mathbf{x})) - S_t(\mathbf{x}) \right), \tag{8}$$

where the same Tukey function is used as ρ_{T_S}. We also tried a quadratic function, but Tukey performed better. Here, we set λ_{T_S} much smaller than λ_{T_A} ($\lambda_{T_A} = 10$ and $\lambda_{T_S} = 1$) so that the shading term has less impact than the albedo term.

3.3 Spatial Shading Prior

Our spatial priors on shading encourage local and non-local smoothness of the estimated shading image. One of our key contributions is to exploit optical flow information in this spatial smoothness prior, resulting in a method that does not require object segmentation or depth data. A similar idea is used to define priors on optical flow in the Classic+NL method [26] and a slightly modified formulation works well for enforcing shading smoothness. Note that optical flow and shading information have some things in common. Both lack the high frequency structure of image texture. Flow and shading are both related to surfaces and change smoothly on smooth surfaces. They also are discontinuous at surface boundaries. These similarities may explain why a spatial smoothness model for flow works well for shading. Our shading term is

$$\begin{aligned} f_S(S_t) = \; & \lambda_{S_s} \sum_{\mathbf{x}} \sum_{\mathbf{y}\in N_3(\mathbf{x})} \rho_S(S_t(\mathbf{x}) - S_t(\mathbf{y})) \\ & + \lambda_{S_m} \sum_{\mathbf{x}} w^{\text{bnd}}_{\mathbf{u}_t}(\mathbf{x}) \cdot \sum_{\mathbf{y}\in N_{\text{nl}}(\mathbf{x})} w^{\text{nl}}_{\mathbf{u}_t}(\mathbf{x},\mathbf{y}) \left| S_t(\mathbf{x}) - S_t(\mathbf{y}) \right|, \end{aligned} \tag{9}$$

where ρ_S is the Charbonnier as above. The weight map $w^{\text{bnd}}_{\mathbf{u}_t}$ is another type of intrinsic image computed from optical flow as shown in Fig. 2 (h); the value is 1 if the pixel is near a motion boundary and 0 otherwise. For this we use a simple

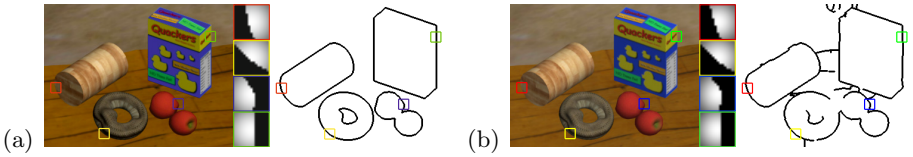

Fig. 3. Examples on the non-local weights defined in Eq. (10). (a) Weights computed from ground truth flow. (b) Weights computed from estimated flow. In each of (a) and (b), small boxes in the middle visualize 15×15 weights corresponding to the regions marked on the left image, and the right image shows motion boundaries detected from optical flow, visualized as $1 - w^{\text{bnd}}_{\mathbf{u}_t}$. Note that the weight function stops spatial propagation around the motion boundaries.

Sobel filter applied to $\mathbf{u}_t$ and dilate the result to obtain the weight map $w^{\text{bnd}}_{\mathbf{u}_t}$. N_3 means a 3×3 window around each pixel to encourage local smoothness, while N_{nl} is a non-local window of 15×15 pixels. Minimizing the non-local term, with the L1 penalty, corresponds to computing a weighted median in the region N_{nl} [21]. Note that we only take the weighted median near the motion boundaries by using $w^{\text{bnd}}_{\mathbf{u}_t}$. While the non-local term improves results at boundaries, applying it everywhere in the image produces over smoothing away from boundaries.

The spatially-varying weight $w^{\text{nl}}_{\mathbf{u}_t}$ encodes information about motion boundaries, which serves as a proxy for surface boundaries. It is defined as follows:

$$w^{\text{nl}}_{\mathbf{u}_t}(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{|\mathbf{x} - \mathbf{y}|^2}{2\sigma_s^2}\right) \cdot \text{surf}_{\mathbf{u}_t}(\mathbf{x}, \mathbf{y}), \tag{10}$$

where $\mathbf{x}$ is the center of a non-local (15×15) window, $\mathbf{y}$ is a pixel in the neighborhood of $\mathbf{x}$, and $\sigma_s = 7$. A binary function $\text{surf}_{\mathbf{u}_t}$ depends on the flow field, $\mathbf{u}_t$, and helps the weight function stop spatial propagation around motion boundaries. It returns 1 at $\mathbf{y}$ if $\mathbf{y}$ and $\mathbf{x}$ stay within the same object region but 0 otherwise: we segment the non-local region into two pieces using the motion boundary inside, and assign 1's to the piece that includes $\mathbf{x}$ and 0's to the other. The boundary weights are illustrated in Fig. 3.

The two weights in Eq. (9) play an important role in preventing over smoothing at motion boundaries (and hence at object boundaries). Note that the weight function used for flow estimation in [26] uses occlusion and color boundaries. For shading, color boundaries are irrelevant and hence we use only motion boundaries. Note that there is no flow for the last frame in the sequence and there we use only the local term; this works well thanks to the information propagated from the previous frames. Other approaches could be used for smoothing with discontinuities; for example, bilateral filtering [22]. In contrast, our approach makes the intrinsic images for occlusions and boundaries explicit.

3.4 Spatial Albedo Prior

To model the spatial variation of albedo we adopt the two relevant spatial priors suggested in [2,4]:

$$f_A(A_t) = \tag{11}$$

$$\lambda_{A_s} \sum_{\mathbf{x}} \sum_{\mathbf{y} \in N_5(\mathbf{x})} \left[-\log \sum_{m=1}^{40} \alpha_m \cdot \mathcal{N}\left(\mathbf{A}_t(\mathbf{x}) - \mathbf{A}_t(\mathbf{y});\ \mathbf{0}, \sigma_m \mathbf{\Sigma}\right) \right]$$

$$-\lambda_{A_p} \log \left[\frac{1}{\mathrm{N}^2 \sqrt{4\pi \cdot \sigma_p^2}} \sum_{\mathbf{x}}^{\mathrm{N}} \sum_{\mathbf{y}}^{\mathrm{N}} \exp\left(-\frac{\|\mathbf{W}(\mathbf{A}_t(\mathbf{x}) - \mathbf{A}_t(\mathbf{y}))\|_2^2}{4\sigma_p^2} \right) \right], \tag{12}$$

where $\mathbf{A}_t(\mathbf{x})$ defines an RGB vector at pixel $\mathbf{x}$, α_m is a mixture constant for each multivariate Gaussian $\mathcal{N}(\cdot)$ whose covariance matrix $\mathbf{\Sigma}$ is scaled by σ_m, $\mathbf{W}$ is a whitening transform matrix to nullify dependency between color channels, σ_p is a standard deviation, N is the number of pixels. The first term shares the concept that underlies Retinex algorithms and encourages small spatial variation of the estimated albedo image based on a multivariate Gaussian scale mixture. The second term models global sparsity of albedo values as proposed in [12]. We use the distribution parameters for these priors learned in [4].

4 Optimization

Traditional approaches optimize for either albedo or shading by assuming that the Lambertian equation (1) is strictly satisfied. However, this assumption does not always hold in practice and thus the solution may be biased to either albedo or shading. Instead, we use the Lambertian equation as a soft constraint (Eq. (3) in Section 3.1) and solve for both variables concurrently. The concurrent optimization is challenging, but our temporal coherence terms effectively constrain the problem. To minimize our objective function, Eq. (2), we adopt a coarse to fine pyramid-based approach and incremental update scheme similar in spirit to the flow estimation method in [26]. Note that the objective function is defined over the entire sequence (not individual frames).

Our new spatial shading prior (Eq. (9) in Section 3.3) is difficult to directly optimize because of the non-local energy term. Instead, an auxiliary "coupling" variable $\tilde{S}_t$ is introduced to assist minimization of the non-local median energy:

$$f'_S(S_t, \tilde{S}_t) = \lambda_{S_s} g_{\mathrm{l}}(S_t) + \lambda_{S_m} g_{\mathrm{nl}}(\tilde{S}_t) + \lambda_{S_{\mathrm{cpl}}} \sum_{\mathbf{x}} w_{\mathbf{u}_t}^{\mathrm{bnd}}(\mathbf{x}) \left\| S_t(\mathbf{x}) - \tilde{S}_t(\mathbf{x}) \right\|^2, \tag{13}$$

where g_{l} and g_{nl} are the local and non-local terms in Eq. (9), respectively. The quadratic term above encourages the estimated S_t and $\tilde{S}_t$ to be the same. We found that $\lambda_{S_{\mathrm{cpl}}} = 10$ works well in our shading estimation problem, with $\lambda_{S_s} = 2$ and $\lambda_{S_m} = 10000$. We alternate between minimizing S_t and $\tilde{S}_t$ as in [26]. More details are given in the supplementary material.

5 Experiments

We evaluate our intrinsic video estimation using three synthetic and three real sequences that illustrate different types of the motion and illumination variation. Due to limited space, we only show the first two frames of two sequences. Our supplementary material[1] includes data generation details and full results, in addition to the optical flow, occlusion and boundary intrinsic images, which are omitted here due to space. The computation time linearly increases with the number of frames (2.2h for 8 frames on average). We use 7 to 9 frame sequences here and, while shorter sequences can be used, we find that the quality improves with more frames because information propagates over all frames.

We compare our results with a baseline color-Retinex algorithm in [13] (CRET) and a more advanced Retinex-based method in [12] (GS). Note that both are *single-image* methods. Existing non-Retinex-based single-image methods only work with additional depth data [4,11] or segmentation of each object [2,3]. Our method deals with a general scene with multiple objects without depth information while making no assumption of rigidity. We obtained CRET results by using the color-Retinex term in the implementation of GS. Our intrinsic video method is denoted "IV". The optical flow used by our method is computed from the input video using Classic+NL [26] .

5.1 Synthetic Examples

Figures 4 and 5 show two synthetic examples with different types of motion and illumination variation. For each sequence we have ground truth values of albedo, shading, optical flow and occlusion. The CRET and GS methods are applied to each frame of the video independently.

As shown in (g)-(r) of the figures, both of CRET and GS put too much high-frequency albedo information into the shading image, and the albedo changes significantly from frame to frame. In contrast, our albedo image retains textural details and the shading is piecewise smooth, mostly obeying object boundaries. Our recovered albedo is consistent over time. One way to see this is by computing the optical flow using the recovered albedo sequences from each method; this is shown in (c)-(f) of the figures. We applied Classic+NL (using brightness instead of texture decomposition), to each reconstructed albedo sequence and the original images. This provides a measure of how temporally coherent the albedo is; an albedo sequence with better temporal coherence will produce flow images that look closer to the ground truth flow ($\mathbf{u}^{\mathrm{GT}}$).

Quantitative Analysis. In Fig. 6 (left), we measure the local mean squared error (LMSE) [13] of the reconstructed albedo and shading images; this is a standard error measure for evaluating intrinsic images. We calculate the LMSE at each frame and average this over all frames, and then average this over all three synthetic examples. We ran IV with both the computed optical flow as

[1] `http://ps.is.tuebingen.mpg.de/project/Intrinsic_Video`

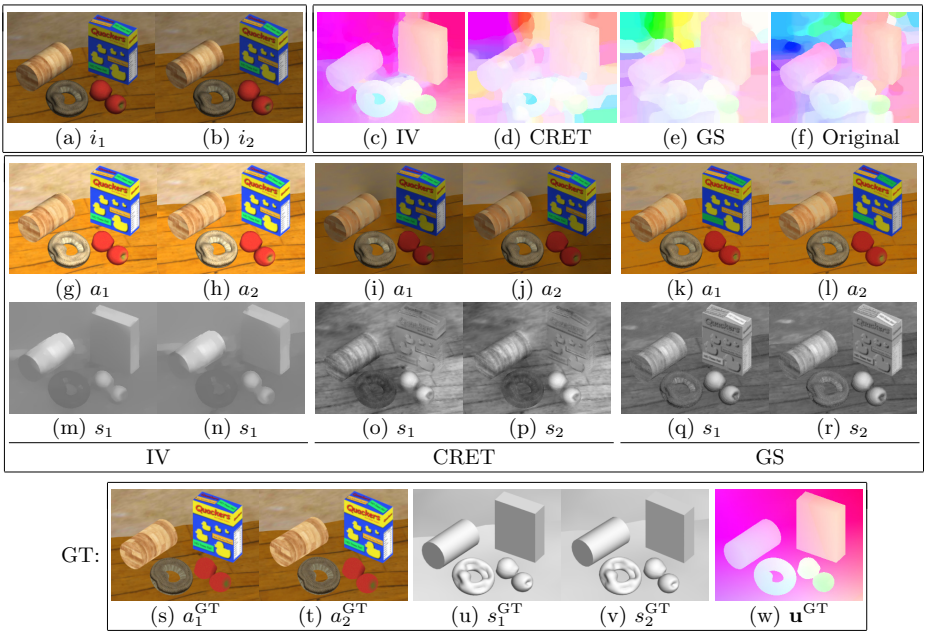

Fig. 4. Synthetic example in which a camera is freely moving and illumination varies significantly over time. (a),(b) Two frames from the sequence. (c)-(e) Flow from the albedo estimated from our method (IV), CRET and GS. (f) Flow from the original images. (g)-(l) Albedo from IV, CRET and GS. (m)-(r) Shading from IV, CRET and GS. (s)-(w) Ground truth albedo, shading and flow.

well as ground truth flow to evaluate the effect of flow errors on the solution. The results are shown in Fig. 6 (left).

Our reconstructed intrinsic images have smaller errors than the GS method: 13.3% with estimated flow and 14.5% with ground truth flow. Note that while ground truth flow improves results slightly, the estimate flow works well. We also disabled the temporal terms (IV w/o flow) to evaluate the importance of motion. In this case we do not use the temporal terms or the motion-based spatial smoothness weighting. More details are given in the supplemental material.

In Fig. 6 (right), we introduce a new temporal incoherence measure that assesses how consistent the reconstructed albedo is over time. Optical flow methods typically assume brightness constancy, which is violated if the illumination is inconsistent over time. Since violations of constancy increase errors in optical flow, the optical flow error provides a measure of how constant an albedo sequence is in time. We compute EPE (averaged end-point-error) [1] of the estimated flow (using estimated albedo sequences) compared with the ground truth flow and then average this over the three synthetic examples. Our albedo sequence is significantly more coherent (lower EPE) than the albedo estimated by previous methods. In addition, note that the flow computed from our albedo is more

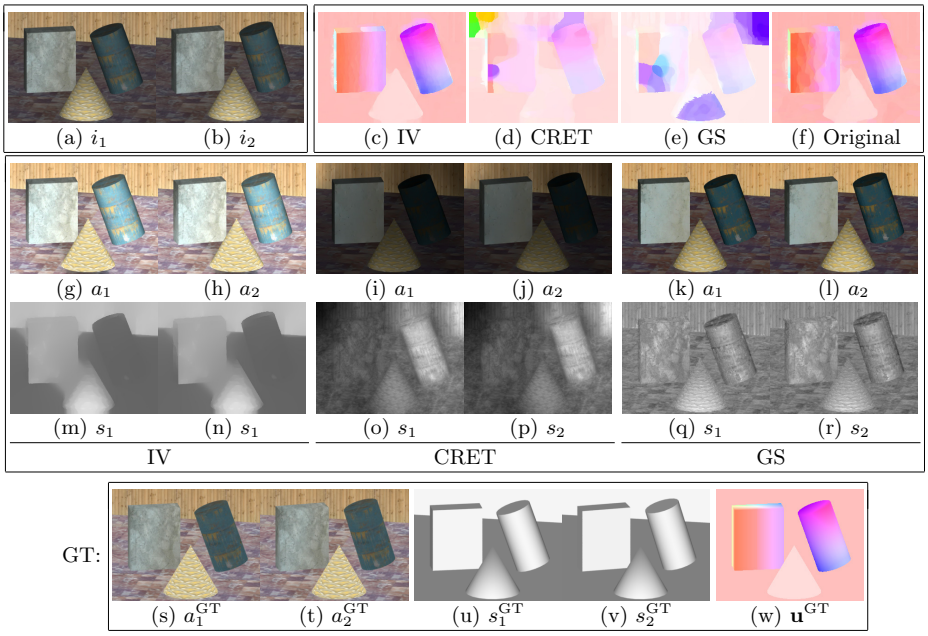

Fig. 5. Synthetic example in which all objects in the scene are moving while the camera translates. Illumination does not change much in this case. (a),(b) Two frames from the sequence. (c)-(e) Flow from the albedo estimated from our method (IV), CRET and GS. (f) Flow from the original images. (g)-(l) Albedo from IV, CRET and GS. (m)-(r) Shading from IV, CRET and GS. (s)-(w) Ground truth albedo, shading and flow.

accurate than the flow computed from the original images. The illumination changes in the original images violate brightness constancy. This result suggests that intrinsic video may be useful to improve optical flow estimation.

5.2 Real Examples

Figures 7 and 8 show two of our real examples. We captured real videos by serially taking photographs with a flashlight or static lighting. The real sequences involve different types of motion and illumination variation, corresponding to those in the synthetic examples. The results are consistent with those on synthetic sequences. As shown in (g)-(r) of the figures, our method significantly outperforms the previous methods. The shading from previous methods carries a lot of albedo information, but our shading sequence has few albedo details and well captures the overall shape of the scene with clean boundaries. The previous methods sometimes almost completely miss the shape of the scene in their shading images and the albedo is overall inconsistent between frames. While there is no ground true flow for this sequence, our reconstructed albedo produces less

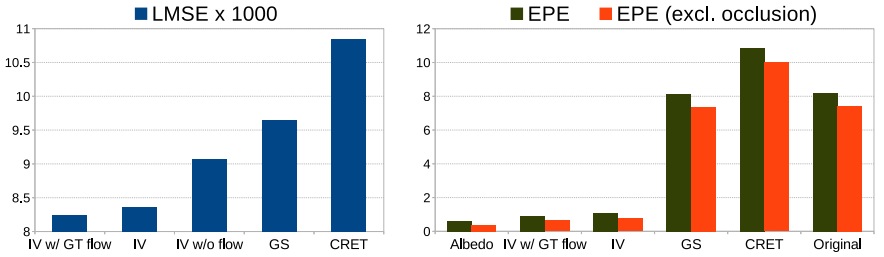

Fig. 6. Quantitative analysis. **Left**: LMSEs of the estimated albedo and shading. Our method produces lower errors than CRET and GS. IV uses estimated flow for the temporal coherence terms. IV performs better than without using the temporal terms (IV w/o flow), and works even better using ground truth flow (IV w/ GT flow). **Right**: EPE (our temporal incoherence measure) of the ground truth albedo sequence (baseline), the albedo sequence estimated by IV (ours), the albedo sequence estimated by GS, the albedo sequence estimated by CRET, and the original video. Our albedo shows better coherence than that from CRET and GS. We measured EPE with and without masking ground truth occlusion areas.

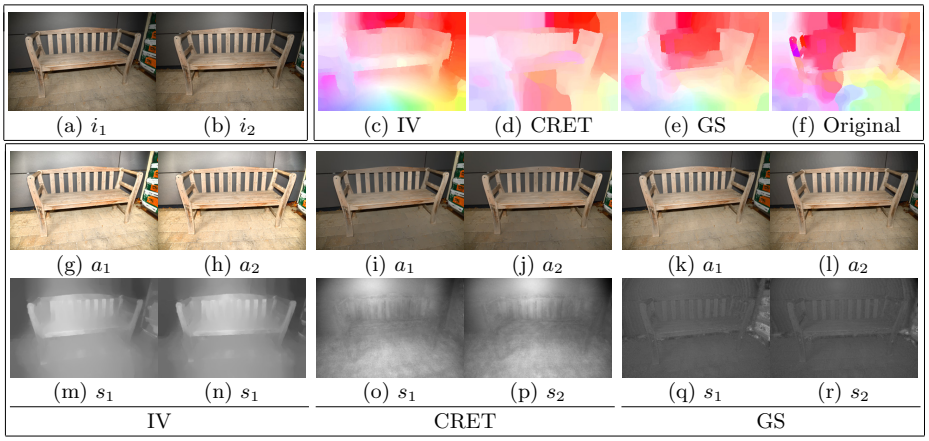

Fig. 7. Real example in which the input video captures a static outdoor scene with a freely moving camera. A flashlight on top of the camera was used to vary illumination over time fairly drastically. (a),(b) Two frames from the sequence. (c)-(e) Flow from the albedo estimated from our method (IV), CRET and GS. (f) Flow from the original images. (g)-(l) Albedo from IV, CRET and GS. (m)-(r) Shading from IV, CRET and GS.

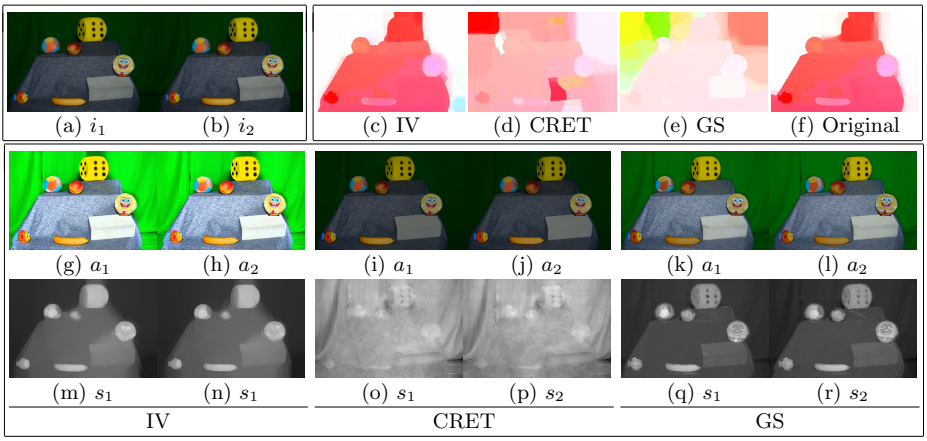

Fig. 8. Real example in which all objects continuously move but the background stays still. The camera and light sources are fixed. (a),(b) Two frames from the sequence. (c)-(e) Flow from the albedo estimated from our method (IV), CRET and GS. (f) Flow from the original images. (g)-(l) Albedo from IV, CRET and GS. (m)-(r) Shading from IV, CRET and GS.

noisy flow fields, suggesting that our albedo has better temporal coherence than the others as illustrated in (c)-(f) of the figures.

6 Conclusions and Future Work

We have introduced the idea of intrinsic video and an algorithm for extracting it automatically from video alone. Experiments with real and synthetic sequences demonstrate that our method generates accurate and temporally coherent albedo and shading, even from videos with non-rigid motion and illumination change. Key to our formulation is the assumption that albedo is mostly constant over time, while shading changes slowly. Optical flow provides the correspondence across time that we exploit to enforce novel temporal constraints on albedo and shading. Our experiments show that these temporal constraints significantly improve albedo and shading estimation. In addition to providing temporal continuity, optical flow gives us information about occlusion and putative surface boundaries; these intrinsic images are important for estimating accurate albedo and spatially coherent shading that is not blurred between objects. As a result of incorporating optical flow, our method works for general scenes, with multiple objects, without need of additional depth data or object segmentation.

According to our incoherence measure, intrinsic video may be useful for optical flow estimation because the resulting albedo sequences obey brightness constancy. Beyond our current work, we believe that integration of the intrinsic video and optical flow problems may produce better results for both. This work provides a new direction for research on both problems. As future work, we will explore the simultaneous estimation of both intrinsic video and optical flow.

References

1. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. International Journal of Computer Vision (IJCV) 92(1), 1–31 (2011)
2. Barron, J.T., Malik, J.: Color constancy, intrinsic images, and shape estimation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 57–70. Springer, Heidelberg (2012)
3. Barron, J.T., Malik, J.: Shape, albedo, and illumination from a single image of an unknown object. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 334–341 (2012)
4. Barron, J.T., Malik, J.: Intrinsic scene properties from a single RGB-D image. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17–24 (2013)
5. Barrow, H.G., Tenenbaum, J.M.: Recovering intrinsic scene characteristics from images. In: Computer Vision Systems, pp. 3–26 (1978)
6. Bell, M., Freeman, W.T.: Learning local evidence for shading and reflectance. In: Proc. IEEE International Conference on Computer Vision (ICCV), pp. 670–677 (2001)
7. Black, M.J., Anandan, P.: The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. Computer Vision and Image Understanding 63(1), 75–104 (1996)
8. Bousseau, A., Paris, S., Durand, F.: User-assisted intrinsic images. ACM Trans. Graphics (TOG) – Proc. SIGGRAPH Asia 28(5), 130:1–130:10 (2009)
9. Bruhn, A., Weickert, J., Schnorr, C.: Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. International Journal of Computer Vision (IJCV) 61, 211–231 (2005)
10. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 611–625. Springer, Heidelberg (2012)
11. Chen, Q., Koltun, V.: A simple model for intrinsic image decomposition with depth cues. In: Proc. IEEE International Conference on Computer Vision (ICCV), pp. 241–248 (2013)
12. Gehler, P., Rother, C., Kiefel, M., Zhang, L., Schölkopf, B.: Recovering intrinsic images with a global sparsity prior on reflectance. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F.C.N., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems (NIPS), pp. 765–773 (2011)
13. Grosse, R., Johnson, M.K., Adelson, E.H., Freeman, W.T.: Ground-truth dataset and baseline evaluations for intrinsic image algorithms. In: Proc. IEEE International Conference on Computer Vision (ICCV), pp. 2335–2342 (2009)
14. Hauagge, D., Wehrwein, S., Bala, K., Snavely, N.: Photometric ambient occlusion. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2515–2522 (2013)
15. Horn, B.K.P., Schunk, B.G.: Determining optical flow. Artificial Intelligence 17, 185–203 (1981)
16. Horn, B.K.P.: Determining lightness from an image. Computer Graphics and Image Processing 3(1), 277–299 (1974)
17. Laffont, P.Y., Bousseau, A., Drettakis, G.: Rich intrinsic image decomposition of outdoor scenes from multiple views. IEEE Transactions on Visualization and Computer Graphics 19(2), 210–224 (2013)

18. Laffont, P.Y., Bousseau, A., Paris, S., Durand, F., Drettakis, G.: Coherent intrinsic images from photo collections. ACM Transactions on Graphics (TOG) – Proc. SIGGRAPH Asia 31(6), 202:1–202:11 (2012)
19. Land, E.H., McCann, J.J.: Lightness and Retinex theory. Journal of the Optical Society of America 61(1), 1–11 (1971)
20. Lee, K.J., Zhao, Q., Tong, X., Gong, M., Izadi, S., Lee, S.U., Tan, P., Lin, S.: Estimation of intrinsic image sequences from image+depth video. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 327–340. Springer, Heidelberg (2012)
21. Li, Y., Osher, S.: A new median formula with applications to PDE based denoising. Communications in Mathematical Sciences 7(3), 741–753 (2009)
22. Paris, S., Kornprobst, P., Tumblin, J., Durand, F.: Bilateral filtering: Theory and applications. Foundations and Trends in Computer Graphics and Vision 4(1), 1–73 (2009)
23. Sand, P., Teller, S.: Particle video: Long-range motion estimation using point trajectories. International Journal of Computer Vision 80(1), 72–91 (2008)
24. Shen, J., Yang, X., Jia, Y., Li, X.: Intrinsic images using optimization. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3481–3487. IEEE (2011)
25. Shen, L., Tan, P., Lin, S.: Intrinsic image decomposition with non-local texture cues. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–7 (2008)
26. Sun, D., Roth, S., Black, M.J.: A quantitative analysis of current practices in optical flow estimation and the principles behind them. International Journal of Computer Vision (IJCV) 106(2), 115–137 (2014)
27. Tappen, M.F., Adelson, E.H., Freeman, W.T.: Estimating intrinsic component images using non-linear regression. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. II: 1992–1999 (2006)
28. Tappen, M.F., Freeman, W.T., Adelson, E.H.: Recovering intrinsic images from a single image. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(9), 1459–1472 (2005)
29. Weiss, Y.: Deriving intrinsic images from image sequences. In: Proc. IEEE International Conference on Computer Vision (ICCV), pp. II: 68–75 (2001)
30. Ye, G., Garces, E., Liu, Y., Dai, Q., Gutierrez, D.: Intrinsic video and applications. ACM Transactions on Graphics 33(4) (2014)