Color naming guided intrinsic image decomposition

Yuanliu Liu, Zejian Yuan*, Member, IEEE

Abstract—Intrinsic image decomposition is a severely under-constrained problem. User interactions can help to reduce the ambiguity of the decomposition considerably. The traditional way of user interaction is to draw scribbles that indicate regions with constant reflectance or shading. However, the effect scopes of the scribbles are quite limited, so dozens of scribbles are often needed to rectify the whole decomposition, which is time consuming. In this paper we propose an efficient way of user interaction that users need only to annotate the color composition of the image. Color composition reveals the global distribution of reflectance, so it can help to adapt the whole decomposition directly. We build a generative model of the process that the albedo of the material produces both the reflectance through imaging and the color labels by color naming. Our model fuses effectively the physical properties of image formation and the top-down information from human color perception. Experimental results show that color naming can improve the performance of intrinsic image decomposition, especially in cleaning the shadows left in reflectance and solving the color constancy problem.

Index Terms—intrinsic images; color naming; user interaction.

1 INTRODUCTION

Intrinsic image decomposition is proposed by [1]. They represent each physical characteristic of the scene by a separate image, and all these intrinsic images are recovered from a single image. The primarily addressed intrinsic images are shading, reflectance, distance, and orientation. Over time, the types of intrinsic images have been limited to reflectance and shading [2], [3], [4]. The reflectance is the albedo of the material, while the shading records the illumination that strikes the surface. The observed image is basically the pointwise product of the reflectance and the shading.

Intrinsic image decomposition is an important preprocessing for many computer graphics and computer vision problems. One important application is image editing. After separating the factors of the scene, users can edit one or more factor while keeping the others untouched. More specifically, recoloring [5], [6], retexturing [7], colorization [8], and makeup simulation [9] change the reflectance, and relighting replaces the illuminants [10], [11]. In terms of computer vision, using the reflectance instead of the raw image for object recognition can increase the robustness to illumination [12]. Reflectance is also used for face alignment [13]. The shading is widely used to recover the 3D shape of the objects by shape-from-shading techniques [14], [15].

Despite the potential of being a valuable preprocessing, intrinsic image decomposition itself is a challenging task that limits its applicability. In essence, intrinsic image decomposition is an underconstrained problem. There are infinite combinations of reflectance and shading that can reproduce the input image. An intuitive solution is to incorporate more constraints to reduce the ambiguity. Usually, the constraints are raised from the prior knowledge of general scenes.

The most widely used priors are the local smoothness of shading [16], [17], [18], [19], the piecewise constancy of reflectance [20], [21], [16], [19], [22], and the global sparsity of reflectance [20], [17], [23], [24], [22]. However, these general priors are not applicable to irregular images or regions. For example, the shading is not smooth at shadow edges. More specific constraints come from the statistics of the illuminants and reflectance. The illuminants and reflectance are supposed to locate at a restricted area or certain peaks in the color space, so the solution space of intrinsic image decomposition will be significantly reduced. These data-dependent constraints can hardly be generalized to other datasets with different statistics.

When reliable prior information is unavailable, we have to seek help from user interactions. The human visual system is able to perceive the intrinsic reflectance of the surfaces by the mechanism of color constancy [27]. Advanced intelligence of discernment, inference, and recognition from the sensation of lights also plays an important role, such as pointed out by Alhazen about a thousand years ago. These abilities enable us to understand the scene thoroughly, so we can infer the intrinsic images in a top-down fashion.

[28] proposed the first user-assisted method of intrinsic image decomposition. They draw scribbles on the image to tell which areas have constant reflectance or shading. These scribbles help solving the ambiguities in determining whether the difference of luminance is caused by reflectance change or shading variation. People are good at finding out those planar areas whose shading change little. They can also locate the shading or shadow edges on surfaces of uniform reflectance according to the geometric or color information. To our best knowledge, all the user-assisted methods of intrinsic image/video decomposition follow this way [29], [6], [10], [30], [31], [32]. The problem is that the constancy of reflectance or shading often appears in very local regions, so quite a lot of scribbles are needed to cover enough portion of the image. It is quite time-consuming to

Y. Liu and Z. Yuan are with the Institute of Artificial Intelligence and Robotics, Xi’an Jiaotong University, Xi’an, Shaanxi, 710049 China. Z. Yuan is the corresponding author.
E-mail: liuyuanliu88@gmail.com, yuan.ze.jian@mail.xjtu.edu.cn

arXiv:1810.09720v1 [cs.CV] 23 Oct 2018
cannot eliminate the effect of shading either. Hence their are unaware of the shading and illumination at all, and the reflectance. The problem is that chip-based methods are then mapped to RGB values that are supposed to be pixels that have the same reflectance. These color names by chip-based method [40], and unify the color names of age decomposition. They obtain the pixel-level color names name descriptors as top-down intervention for intrinsic image. There are also some early attempts to use color reflectance gave more accurate color labels than on the raw prior work [38] demonstrated that color naming on the reflectance has already been addressed in the literature. Our The relation between the color composition and the reflectance is not perfect [36]. Human vision also shows some degree of color constancy [37]. Therefore the annotation of the color composition will not be affected much by the shading or the illumination.

We represent the color composition by a distribution over the eleven basic color terms [33]. In English they are black, blue, brown, gray, green, orange, pink, purple, red, white, and yellow. People are familiar with these concepts, no matter what language they are speaking, so they can describe the color composition quickly and consistently.

We built a generative model that connects the intrinsic images and the color composition through the albedo of materials. The underlying assumption is that the albedo of materials generates both the reflectance in imaging and the color composition in color naming. Based on the global sparsity of reflectance [34], [22], we model the albedo of the materials in the scene by a Gaussian Mixture Model, and the intrinsic reflectance can be regarded as an observation of the albedo. There are also strong evidences for the close relation between the color composition and the reflectance. The physical attribute that primarily determines the perceived color of a surface is exactly its reflectance [27]. Humans are good at intrinsic image decomposition [35], although not perfect [36]. Human vision also shows some degree of color constancy [37]. Therefore the annotation of the color composition will not be affected much by the shading or the illumination.

1.1 Related Work

The relation between the color composition and the reflectance has already been addressed in the literature. Our prior work [38] demonstrated that color naming on the reflectance gave more accurate color labels than on the raw image. There are also some early attempts to use color names for intrinsic image decomposition. [39] take color-name descriptors as top-down intervention for intrinsic image decomposition. They obtain the pixel-level color names by chip-based method [40], and unify the color names of pixels that have the same reflectance. These color names are then mapped to RGB values that are supposed to be the reflectance. The problem is that chip-based methods are unaware of the shading and illumination at all, and unifying the color names of pixels with the same reflectance cannot eliminate the effect of shading either. Hence their color composition does not necessarily correspond to the reflectance. In comparison, human annotation of color composition can reduce the influence of shading and illumination considerably. [12] built a unified framework to jointly estimate the intrinsic images, objects and attributes (including color names). They verified that utilizing the correlations between intrinsic images and attributes do improve the performance of sub-tasks.

Our model is based on the model of automatic intrinsic image decomposition proposed by [17]. Their model is written in the form of Conditional Random Field, taking the smoothness of shading, the sparsity of reflectance, and the Retinex term as additive energy terms. Apart from these common terms, our model differs from the model of [17] in the following aspects:

- Our model utilized the top-down information from color naming to guide the clustering of reflectance, while their clustering are totally data-driven;
- We addressed the direct and ambient illuminations explicitly, and solved the problem of color constancy;
- We use Gaussian mixture model instead of k-means to represent the clusters of reflectance, which facilitates the fusion of color naming guidance and data mode in a probabilistic framework.

Recently, [41] proposed a model based on Dirichlet process Gaussian mixture model. In this model the image was treated as an observation from a generative, stochastic process, and the shading was modeled by a Gaussian process. One superiority of this model is that the optimal number of components for the reflectance image can be decided automatically. However the complexity of this model makes...
2 Generative Model of Image and Color Composition

We build a generative model of the image $I$ and its color composition $y$ as well as the intrinsic images, including the shading $S$, the reflectance $R$, and the illumination $L$. The sketch of our model is depicted in Fig. 2. The core of our model is the albedo of the materials $\theta$ in the scene, which generates not only the reflectance $R$ but also the color composition $y$ of the image in a color naming process.

Given the image and the color composition, the unobservable variables in our model can be determined by solving a MAP problem, where the posterior probability is defined as follows:

$$P(S, R, \theta, L|I, y) \propto P(L)P(S)P(\theta)P(y|\theta)P(R|\theta)P(I|S, R, L).$$ (1)

The probabilities $P(L)$, $P(S)$, and $P(\theta)$ are the priors on the illumination, the shading, and the albedo of materials, respectively. The conditional probabilities $P(y|\theta)$ and $P(R|\theta)$ are the likelihoods of observing the color composition $y$ and the reflectance $R$ on materials $\theta$, respectively. The likelihood of the image $P(I|S, R, L)$ is often expressed in a deterministic way, i.e., $I = R \cdot L \cdot S$.

Our model has two key features. Firstly, the color composition is involved in our model. The image-level color composition affords informative cues on the colors of individual materials together with their mixture proportions, which will further shape the overall distribution of the reflectance.

Secondly, the global illumination is separated from the reflectance and the shading. Traditionally, the illumination is either combined into the shading [28, 41, 4] or ignored [17, 20] showed that separating the illumination can make the decomposition more complete and meaningful. In our former work [42] the color constancy problem was avoided by defining the reflectance to be the fully lighted image that all the regions are covered by full direct illumination. In this case the reflectance is actually the illumination-modulated reflectance $\hat{R} = R \cdot L$. Here we have to separate the illumination from the reflectance, since the perceived color compositions are actually describing the reflectance while the illumination should have been “whitened” by the human visual system.

2.1 Representations of Major Components

The color composition is defined to be a vector $y = [y_1, ..., y_M]^T$, each dimension of which denotes the proportion of one color within the whole image. Specifically, we choose the colors from eleven basic color terms [33], so $M = 11$. Note that we always have $||y||_1 = 1$.

Following our former work [42], we represent the raw images and the intrinsic images in the $UVB$ color space. This color space is formed by a 2D shadow-free plane $UV$ [43] and a brightness dimension $B$ that depends on the shading. The image in the $UVB$ space is obtained from rotating the log RGB space as follows:

$$[I^u_p, I^v_p, I^b_p] = \ln(I_p)H,$$ (2)

where $p$ is the index of pixel. The rotation matrix $H = [u, v, n]$ is mainly determined by the brightening direction $n$, while $u$ and $v$ are merely an arbitrary pair of basis vectors of the $UV$ plane perpendicular to $n$.

Denote the illumination-modulated reflectance in the $UVB$ space by $R_p = [R^u_p, R^v_p, R^b_p]$. According to the property of shadow-free plane $UV$ [43], we have the following approximations:

$$R^u_p \approx I^u_p = \ln(I_p) \cdot u$$
$$R^v_p \approx I^v_p = \ln(I_p) \cdot v.$$ (3)

The only unknown dimension of the illumination-modulated reflectance is the reflectance brightness $R^b_p$, which is the main goal of our inference in Section 3. Once we obtain $R^b_p$, we can recover the illumination-modulated reflectance in the RGB space by:

$$\hat{R}_p = exp([R^u_p, R^v_p, R^b_p]H^{-1}),$$ (4)

where $exp$ denotes element-wise exponential.

According to the definition of the illumination-modulated reflectance, the reflectance can be calculated by $R_p = \hat{R}_p/L$. In the $UVB$ color space the reflectance will be:

$$\hat{R}_p = [R^u_p, R^v_p, R^b_p] - [L^u, L^v, L^b]$$
$$= [I^u_p - L^u, I^v_p - L^v, R^b_p - L^b],$$ (5)

where $[L^u, L^v, L^b]$ is the illumination in the $UVB$ space. Here we utilized the approximations of $R^u$ and $R^v$ in Equation (3). Note that the illumination in the $UVB$ space acts as a global bias, which does not affect the shape of the distribution of pixels.

Most natural images can be represented by a limited number of color lines [34], each of which corresponding to a material with unique color. After removing the variance of brightness caused by shading, the reflectance of pixels falls naturally into a limited number of compact clusters in the $UVB$ color space as shown in Fig. 3. Therefore we build a Gaussian mixture model (GMM) for the albedo of materials $\theta$ in the $UVB$ color space. The GMM of $\theta$ is specified by the means $\mu$, the covariances $\Sigma$, and the mixing coefficients $\pi$. Each Gaussian component is supposed to cover the pixels from a certain material, whose albedo, covariance of color, and population are recorded in the corresponding component of $\mu$, $\Sigma$ and $\pi$, respectively.

Discussion. The $UVB$ color space addresses the bi-illumination scenario [43], where there is a direct illuminant.
The MAP problem in Equation (1) is transformed into an energy minimization problem as usual. The posterior probability is realized by the following energy function:

\[ E(R^b; \theta, \mathbf{L}; \mathbf{l}, \mathbf{y}) = w_s E_s(R^b) + w_r E_r(R^b) + w_g E_g(\theta) + E_d(R^b, \theta, \mathbf{L}; \mathbf{l}) + w_c E_c(\theta; \mathbf{y}) \]

\[ \text{s.t.} \quad \mathbf{L} \in \mathcal{L}, \quad ||\pi||_1 = 1, \quad \pi_k \geq 0, \forall k, \]

where \( w_s, w_r, w_g \) and \( w_c \) are the weights of the energy terms. The illumination \( \mathbf{L} \) is assumed to be uniformly distributed within the feasible domain \( \mathcal{L} \), so \( P(\mathbf{L}) \) in Equation (1) can be omitted. The optimal illumination is selected to be the one that makes the color composition of reflectance consistent with the annotation. The energy terms are described in the following paragraphs.

The smoothness of shading \( E_s \) and the Retinex term \( E_r \) are similar to those used by [17] but formulated in the \( UVB \) space:

\[ E_s(R^b) = \sum_{p \sim q} (R^b_p - R^b_q - (I^b_p - I^b_q))^2, \]  

and

\[ E_r(R^b) = \sum_{p \sim q} (R^b_p - R^b_q - g_{p,q}(\mathbf{l}) \cdot (I^b_p - I^b_q))^2, \]

where \( p \sim q \) denotes the neighborhood relation in a 4-connected pixel graph. \( g_{p,q}(\mathbf{l}) \in \{1, 0\} \) indicates whether the edge between pixels \( p \) and \( q \) is a reflectance edge or not. Following Color Retinex [2] we calculate the gradients in chromaticity (\( UV \) dimensions) and brightness. If the magnitude of the gradient of chromaticity is greater than a threshold \( T^c \) or the gradient of brightness is greater than a threshold \( T^b \), we classify the edge to be a reflectance edge. The thresholds are chosen by cross-validation. The shading brightness \( S^b \) is replaced by \( I^b - R^b - L^b \). If there is no reflectance edge between pixel \( p \) and its neighbor \( q \), the shading order \( S^b_p - S^b_q \) should be equal to their brightness order \( I^b_p - I^b_q \). Otherwise the shading order is expected to be around 0 due to the smoothness of shading. Note that \( E_s \) is a special case of \( E_r \) when \( g_{p,q}(\mathbf{l}) = 1 \). Nevertheless \( E_s \) is still useful to suppress the sharp shading changes where the reflectance edges are not detected out.

The prior of the global sparsity of reflectance is realized by minimizing the variances of the clusters:

\[ E_g(\theta) = \sum_k \ln |\Sigma_k|, \]

where \( | \cdot | \) is the determinant of matrix. Note that, the covariance matrices here are specified to be diagonal, and they are always positive definite.

The reflectance term is represented by the negative log likelihood of reflectance under the GMM model:

\[ E_d(R^b, \theta, \mathbf{L}; \mathbf{l}) = -\ln P(\hat{\mathbf{R}} | \theta) = -\sum_p \ln \left( \sum_k \pi_k \mathcal{N}(\hat{\mathbf{R}}_p | \mu_k, \Sigma_k) \right), \]

where the reflectance \( \mathbf{R} \) in the \( UVB \) space is calculated by Equation (5). To decouple the variables of different components, a binary matrix \( z = [z_{p,k}]_{(p,k)=(1,1)} \) is introduced to denote the assignments of the \( N \) pixels to the \( K \) clusters [45]. Define the probability of assignment as \( \gamma(z_{p,k}) = P(z_{p,k} = 1 | \hat{\mathbf{R}}_p) \). The reflectance term can be rewritten to be:

\[ E_d(R^b, \theta, \mathbf{L}; \mathbf{l}) = -\sum_{p,k} \gamma(z_{p,k}) \{ \ln \pi_k + \ln \mathcal{N}(\hat{\mathbf{R}}_p | \mu_k, \Sigma_k) \}. \]

The color naming term is modeled by:

\[ E_c(\theta; \mathbf{y}) = ||\mathbf{y} - \hat{\gamma}(\mu)\pi||^2_2, \]

where \( \hat{\gamma}(\mu) = [\hat{\gamma}(\mu_1), ..., \hat{\gamma}(\mu_K)] \) is the component-level color composition. The color composition of the \( k \)-th component of the GMM is computed by \( \hat{\gamma}(\mu_k) = T(\exp(\mu_k \cdot \hat{H}^{-1})) \), where \( T \) is the color naming projection from the RGB space to the color name space. In our experiments we adopt the method of [40], which is a well-known chip-based color naming method. Here we apply color naming on the cluster centers instead of individual pixels, which saves a lot of computation time.

1. The gradient of brightness at reflectance edges are often not as large as deep shadow edges [27].
of computation cost. Color naming on pixel clusters seems more likely to be the way that people determine the color composition of images. The color assimilation effect will promote the perceived color of clustered pixels to be the same [46]. Therefore the variances of colors within the clusters are unlikely to change the color naming results. Here the image-level color composition of the reflectance is the weighted sum of the component-level color compositions, while the weights are the mixing coefficients of the GMM.

**Discussions.** The color naming term is the key to realize color naming guided intrinsic image decomposition. Minimizing the energy $E_c$ with respect to $\pi$ will adapt the mixture coefficients of the GMM until the accumulation of component-level color compositions approximates the image-level color composition. After that, the assignments of the pixels to the components will be updated in a way that the larger the mixture coefficient is, the larger the population of the corresponding Gaussian component will be. Especially, the components formed by shadowed pixels will be gradually merged into the components of normally lighted pixels, and their brightness will be raised accordingly during the minimization of the reflectance term $E_d$ in Equation (11).

### 3 Inference

We minimize the energy function in Equation (6) in an iterative way. The whole process is summarized in Algorithm 1. The key operation is to search the optimal mixing coefficients $\pi$, which not only explain the distribution of the components in the GMM but also reconstruct the image-level color composition from the component-level color compositions. The temporal results of an example image in the iterative procedure are shown in Fig. 4.

**Optimize $R^b$ by:**

$$R^b_{t+1} = \arg \min_{R^b} w_s E_s(R^b) + w_r E_r(R^b) + E_d(R^b, \theta_t, L_t; I).$$  

(13)

All the energy terms are differentiable, so the optimization can be solved efficiently by gradient descend [47].

**Select illumination $L$ by:**

$$L_{t+1} = \arg \min_{L \in \mathcal{L}} E_d(R^b_{t+1}, \theta_t, L; I).$$  

(14)

The feasible domain of illumination $L$ covers the chromaticity of all the illumination of the dataset, while the intensity is empirically set to be within $[0, 2]$. The feasible domain is then sampled uniformly, and the optimal illumination is selected from those samples.

**Update GMM $\theta$.** We minimize the energy $E(R^b, \theta, L; I, y)$ in Equation (6) with respect to the parameters $\theta$ of GMM by the EM algorithm. In the E step, we evaluate the probability of the assignments of the $N$ pixels to the $K$ clusters by:

$$\gamma(z_{p,k}) = \frac{\pi_k \mathcal{N}(\hat{R}_p; \mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(\hat{R}_p; \mu_j, \Sigma_j)}.$$  

(15)

In the M step, we first update the mean $\mu$ and the variance $\Sigma$ as follows:

$$\mu_k = \frac{1}{N_k} \sum_p \gamma(z_{p,k}) \hat{R}_p$$

$$\Sigma_k = \frac{1}{N_k + 2w_g} \sum_p \gamma(z_{p,k})(\hat{R}_p - \mu_k)(\hat{R}_p - \mu_k)^T,$$

(16)

where $N_k = \sum_p \gamma(z_{p,k})$ is the population of the $k$-th component. It should be pointed out that we do NOT minimize $E_c$ with respect to $\mu$, since the color naming function such as that defined by [40] is too complex to be optimized analytically. Through experiments we found that this simplification did not change the results much. The change of $\mu$ in a single iteration is usually too small to affect color naming. Note that the formulation of $\Sigma_k$ in Equation (16) is slightly different from those in the general EM algorithm for fitting GMM [45], with an extra term $w_g$ in the denominator that reduces the variance. The reason is that the sparsity of reflectance (defined as $E_g$ in Equation (9)) is incorporated into the energy function.

The optimal mixture coefficient $\pi$ is obtained by:

$$\arg \min_{\pi} -N^T \ln \pi + w_c \| y - \hat{y}(\mu) \|_2^2$$

s.t. $\| \pi \|_1 = 1, \pi_k \geq 0, \forall k.$

(17)

We solve this problem by the Alternating Direction Method of Multipliers (ADMM) [48]. We reformulate the problem in Equation (17) as follows:

$$\arg \min_{\pi, \phi, \psi} f(\pi) + g(\phi) + h(\psi)$$

s.t. $\| \pi \|_1 = 1$

$$\pi = \phi = \psi,$$

(18)

where

$$f(\pi) = w_c \| y - \hat{y}(\mu) \|_2^2$$

$$g(\phi) = -N^T \ln \phi$$

$$h(\psi) = \begin{cases} 0 & \text{if } \psi_k \geq 0, \forall k \\ \infty & \text{otherwise.} \end{cases}$$

(19)
Through introducing Lagrange multipliers $\lambda$, $\Gamma$, and $\Upsilon$, we can obtain the following augmented Lagrangian [48]:

$$
L(\pi, \phi, \psi, \lambda, \Gamma, \Upsilon) = f(\pi) + g(\phi) + h(\psi) + \lambda(\|\pi\|_1 - 1) + \Gamma^T(\pi - \phi) + \frac{\rho}{2}\|\pi - \phi\|_2^2 + \Upsilon^T(\pi - \psi) + \frac{\rho}{2}\|\pi - \psi\|_2^2,
$$

where $\rho$ is a scaling parameter, which is empirically set to be 20. We initialize $\pi$, $\phi$, $\psi$, $\Gamma$, and $\Upsilon$ with $\pi_t$, while $\lambda = 1$. Then we update them iteratively as follows:

$$
\pi_{t+1} = \pi_t - \eta \left(2w_s (y(\mu)^T \pi_t - 2w_s y(\mu)y + \Gamma_t + \rho(\pi_t - \phi_t) + \Upsilon_t + \rho(\pi_t - \psi_t) + \lambda_t) + \phi_{t+1} = \phi_t - \eta \left(- (\text{diag}(\pi_{t+1}))^{-1} N - \Gamma_t - \rho(\pi_t - \phi_t) \right) \right) \\
\psi_{t+1} = (\pi_{t+1} + 1) \rho \Upsilon_t + \theta (\pi_{t+1} + \Upsilon_t) + \lambda_{t+1} = \lambda_t + \eta(\|\pi_{t+1}\|_1 - 1) \\
\Gamma_{t+1} = \Gamma_t + \eta(\pi_{t+1} - \phi_{t+1}) \Upsilon_{t+1} = \Upsilon_t + \eta(\pi_{t+1} - \psi_{t+1}),
$$

where $(\cdot)^+$ truncates all the elements of a vector to be non-negative. $\eta$ is the step size, which is set to be 0.001 in our experiments. We terminate the iteration when the reduction of the objective function in Equation (17) is less than a threshold $T_d$, and we set $T_d$ to be $10^{-6}$ in implementation. The $\psi$ of the last round is taken to be the output $\pi_{t+1}$.

**Initialize $R^b$.** We initialize the brightness of reflectance by $R^b = I^b - \hat{S}^b$, where $\hat{S}^b$ is the shifted shading brightness got from compensating the categorical bias of the shading brightness, as described by [42]. Especially, the number of categories is set to be the number of positive elements in the color composition $y$. Since the brightness is ignored in the clustering process of the initialization stage, pixels with the same chromaticity but different intensities will be put into the same category. These pixels will be separated into different GMM components in the iterative updates when the other cues such as the smoothness of shading are incorporated.

**4 Experiments**

We evaluate our method on the MIT intrinsic image dataset [2], which is a widely used benchmark for intrinsic image decomposition. This dataset contains challenging images with deep shadows. The results are measured by different types of metrics including MSE, LMSE [2], aLMSE and correlation [49]. We also test on the images supplied by [28], which are commonly used by user-assisted methods. The color compositions of all the test images are annotated by the authors of this paper, each of which takes only a couple of seconds to annotate.

In our experiments the number of components $K$ in the initial GMM is set to be 25. The weights of energy terms $w_s$,
TABLE 1: Results on the MIT Intrinsic Images dataset. Higher correlation, lower MSE, LMSE and aLMSE are better.

|                   | Correlation | MSE   | LMSE  | aLMSE |
|-------------------|-------------|-------|-------|-------|
| Gehler et al.     | 0.7748      | 0.0985| 0.0244| 0.2544|
| Chang et al.      | -           | -     | 0.0229| -     |
| Serra et al.      | 0.7862      | 0.0834| 0.0340| 0.2958|
| IID w/o CN        | 0.8378      | 0.0719| 0.0209| 0.2221|
| CN-IID            | 0.8489      | 0.0687| 0.0204| 0.2152|
| GtCN-IID          | 0.8494      | 0.0676| 0.0203| 0.2129|

$w_r$, $w_g$, and $w_c$ in Equation (6) are set to be 10, 100, 0.5$N/K$, and 5$N$, respectively. Here $N$ is the number of pixels. The threshold of energy drop $\delta$ in Algorithm 1 is set to be 0.01.

To process the images in the Intrinsic image dataset, our algorithm takes about eight minutes for each image on average.

4.1 The Effects of Color Naming Guidance

Fig. 5 illustrates how the color naming guidance changes the results of intrinsic image decomposition. For comparison we execute a version of our method that the color naming term $E_c$ is excluded from the energy in Equation (6). The most significant improvement brought by color naming guidance is that the overall color composition of the recovered reflectance gets much closer to our color perception. For example, the raccoon in the top left part of Fig. 5 has a white body, and color naming guided intrinsic image decomposition (CN-IID) outputs the reflectance in the right color. Without color naming guidance, the recovered reflectance is darker than the groundtruth and the body of the raccoon appears to be gray. The performance on local areas has also been improved notably. For instance, the cup in the bottom left part of Fig. 5 has very different shading between the left side and the right side of the image. The color composition plays two important roles in dealing with the deep shadows: (1) cleaning the shadow residual in the reflectance by eliminating the color of brown that does not occur in the color labels; and (2) choosing a proper number of categories for initializing the reflectance brightness $R_b$ (Section 3). As a by-product CN-IID can produce pixel-level color naming results after recovering the reflectance. This is a promising way of illumination-robust color naming.

Fig. 6 demonstrates the reflectance recovered under different color naming guidance. The user labels the shell of the turtle to be orange (the second column of Fig. 6), which drives the reflectance to be slightly brighter than the groundtruth. When taking the color composition of the groundtruth reflectance as input (the third column of Fig. 6), the shell of the turtle appears to be brown. Both of these outputs are reasonable, and the shadows in local regions are mostly removed from the reflectance. When the input color composition goes wrong, the reflectance will be ruined. For example, in the last column of Fig. 6 the color of “black” occupies too much proportion of the color composition, so the shadows are left in the reflectance to produce black colors.

The quantitative results under different color naming guidance are given in Table 1. Generally, the differences are small, since all the metrics are invariant to the global scale of the intensity and the chromaticity of the intrinsic images, which are the major concerns of adopting color naming guidance. Nevertheless, the quantitative results re-
Comparison, CN-IID needs to input only eleven numbers.

Drawing these scribbles is quite time consuming. In general, dozens of strokes are required for obtaining good results, as shown in Fig. 8. They draw a lot of scribbles to tell which areas have constant reflectance or shading. Typically, dozens of strokes are needed when the global structure has already been specified. It is expected that only a few scribbles are needed when the global structure has already been specified by color composition.


draw a lot of scribbles to tell which areas have constant reflectance or shading. Typically, dozens of strokes are required for obtaining good results, as shown in Fig. 9. Drawing these scribbles is quite time consuming. In comparison, CN-IID needs to input only eleven numbers (the percentages of basic color terms), and the results are comparable to the results of [28]. Since humans are quite familiar with basic color terms, they can give a rough annotation quickly, often within a few seconds. For complex images such as those in the bottom row of Fig. 9, the precise color compositions are unavailable. Nevertheless, rough color compositions work well enough in these cases.

Image-level color composition cannot determine the reflectance precisely, however. Several typical failure cases are shown in Fig. 10. One important defect of CN-IID is the intra-category variation of color names: Each color category covers a large area of the color space [39], so the precision of the color of the recovered reflectance is not high enough. For example, in Fig. 10(a), the pillow in the red box appears blue in the recovered reflectance, but not exactly the same blue that we perceived from the image. Another flaw is the block effect that can be observed in the sky of Fig. 10(b). Textures and smooth change of color are removed when the reflectance of each cluster is unified. This property is shared by most cluster-based methods [35], [41], [17], [23], [19], [39]. Moreover, the minor categories, e.g., white and black in Fig. 10(c), may possibly be merged into other categories.

5 Conclusions and Discussions

We have presented a new method of user-assisted intrinsic image decomposition that takes color composition as input. The color composition afforded informative constraints on the overall distribution of reflectance, which made intrinsic image decomposition more solvable. In our experiments, the performance was improved considerably over the baseline, even when only a rough color composition was available. Compared to the traditional scribble-based user interaction, color naming is efficient yet effective.

We built a generative model to combine the top-down guidance and the bottom-up evidence together. These two sources of information were fused in a probabilistic framework, which give the best explanation of both user perception and image features. However, the features we used here are quite simple, which may not be discriminative enough for all kinds of scenes. Utilizing more powerful features, such as those learned from deep learning [50], might further improve the results.

The image-level color composition has a low resolution, so it cannot determine the reflectance precisely. To overcome this limitation, we consider to incorporate the scribbles into CN-IID in our future work. In this way both the global distribution and the details of intrinsic images can be recovered precisely. It is expected that only a few scribbles are needed when the global structure has already been specified by color composition.

Illumination-robust color naming is a challenging task [38]. Humans may also make mistakes in this task, especially when they are misled by color illusions. In this case color guidance may distract intrinsic image decomposition. We should better discard the color naming guidance if we are not sure about what the real colors are.

Acknowledgment

This work was supported by National Basic Research Program of China (973 Program) [No. 2015CB351703] and [No.
Fig. 7: Comparison to automatic intrinsic image decomposition. The first column shows the input image (top) and the annotation of color composition (bottom) that is only used by the method of CN-IID. For the other columns, the rows from top to bottom are the output shading, reflectance, and the color composition of reflectance, respectively. The output color labels is produced by mapping the RGB values of reflectance to basic color terms using the method of parametric fuzzy sets [40]. Clearly, the reflectance obtained by CN-IID is visually the most close one to the groundtruth.

Fig. 8: Typical results of the MIT intrinsic image dataset.

REFERENCES

[1] H. G. Barrow and J. M. Tenenbaum, Recovering Intrinsic Scene Characteristics from Images, ser. Computer Vision Systems. Academic Press, 1978.

[2] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman, “Ground-truth dataset and baseline evaluations for intrinsic image algorithms,” in IEEE International Conference on Computer Vision, 2009.

[3] B. K. Horn, “Determining lightness from an image,” Computer Graphics and Image Processing, vol. 3, no. 4, pp. 277 – 299, 1974. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0146664X74900227

[4] M. F. Tappen, W. T. Freeman, and E. H. Adelson, “Recovering intrinsic images from a single image,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 9, 2005.

[5] S. Beigpour and J. van de Weijer, “Object recoloring based on intrinsic image estimation,” in IEEE International Conference on Computer Vision, 2011.

[6] R. Carroll, R. Ramamoorthi, and M. Agrawala, “Illumination decomposition for material recoloring with consistent interreflections,” ACM Transactions on Graphics, vol. 30, no. 4, p. 43, 2011.

[7] Y. Xing, S. Jianbing, H. Ying, and M. Xiaoyang, “Re-texturing by intrinsic video,” in International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2010, pp. 486–491.

[8] X. Liu, L. Wan, Y. Qu, T.-T. Wong, S. Lin, C.-S. Leung, and P.-A. Heng, “Intrinsic colorization,” ACM Transactions on Graphics, vol. 27, no. 5, pp. 1–9, 2008.

[9] C. Li, K. Zhou, and S. Lin, “Simulating makeup through physics-based manipulation of intrinsic image layers,” in IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4621–4629.

[10] Y. Dong, X. Tong, F. Pellacini, and B. Guo, “Appgen: interactive material modeling from a single image,” ACM Transactions on Graphics, vol. 30, no. 6, pp. 1–10, 2011.

[11] Z. Liao, K. Karsch, and D. Forsyth, “An approximate shading model for object relighting,” in IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5307–5314.

[12] V. Vineet, C. Rother, and P. H. S. Torr, “Higher order priors for joint intrinsic image, objects, and attributes estimation,” in Advances in Neural Information Processing Systems, 2013, pp. 557–565.

[13] H. Yuchi, S. Lin, L. Hanqing, and S. Heung-Yeung, “Face alignment using intrinsic information,” in International Conference on Image Processing, vol. 5, 2004, pp. 3307–3310.

[14] B. K. P. Horn, Obtaining shape from shading information. MIT Press, 1989, pp. 123–171.

[15] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, “Shape from shad-
Fig. 9: Comparison to scribble-based user interaction. The user scribbles are used by the method of [28]. White scribbles indicate fully-lit pixels. Blue scribbles correspond to pixels sharing a similar reflectance. Red scribbles correspond to pixels sharing a similar illumination. The color compositions are used by CN-IID.
Fig. 10: Failure cases. The red boxes indicate typical regions where CN-IID failed to recover the reflectance. (a) Intra-category variance of color names makes the recovered reflectance deviate from the perceived color. (b) Texture are erased by cluster-based method. (c) Minor categories are sacrificed.

[35] S. Bell, K. Bala, and N. Snavely, “Intrinsic images in the wild,” ACM Transactions on Graphics, vol. 33, no. 4, 2014.
[36] E. H. Adelson, “Lightness perception and lightness illusions,” in In: The New Cognitive Neurosciences, 2nd ed., M. Gazzaniga, Ed.
Cambridge, MA: MIT Press, 2000, pp. 339–351.
[37] K. T. Blackwell and G. Buchsbaum, “Quantitative studies of color constancy,” Journal of The Optical Society of America A, vol. 5, no. 10, 1988.
[38] Y. Liu, Z. Yuan, B. Chen, J. Xue, and N. Zheng, “Illumination robust color naming via label propagation,” in IEEE International Conference on Computer Vision, 2015, pp. 621–629.
[39] M. Serra, O. Penacchio, R. Benavente, and M. Vanrell, “Names and shades of color for intrinsic image estimation,” in IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 278–285.
[40] R. Benavente, M. Vanrell, and R. Baldrich, “Parametric fuzzy sets for automatic color naming,” Journal of The Optical Society of America A, vol. 25, no. 10, 2008.
[41] J. Chang, R. Cabezas, and J. W. Fisher III, “Bayesian nonparametric intrinsic image decomposition,” in Proceedings of the European Conference on Computer Vision, 2014.
[42] Y. Liu, Z. Yuan, and N. Zheng, “Intrinsic image decomposition from pair-wise shading ordering,” in Asian Conference on Computer Vision, 2014, pp. 83–98.
[43] B. Maxwell, R. Friedhoff, and C. Smith, “A bi-illuminant dichromatic reflection model for understanding images,” in IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[44] C. Che-Han, C. Yu-Ting, and C. Yung-Yu, “A non-local sparse model for intrinsic images,” in Asian Conference on Pattern Recognition, 2013, pp. 100–104.
[45] C. Bishop, Pattern Recognition and Machine Learning. Springer, 2007.
[46] B. R. Conway, “Color vision, cones, and color-coding in the cortex,” The Neuroscientist, vol. 15, no. 8, 2009.
[47] C. E. Rasmussen, “Minimize a multivariate function.” [Online]. Available: www.gatsby.ucl.ac.uk/~edward/code/minimize
[48] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” Foundation and Trends in Machine Learning, vol. 3, no. 1, pp. 1–122, 2011.
[49] X. Jiang, A. Schofield, and J. Wyatt, “Correlation-based intrinsic image extraction from a single image,” in Proceedings of the European Conference on Computer Vision, 2010.
[50] T. Zhou, P. Krahenbuhl, and A. Efros, “Learning data-driven reflectance priors for intrinsic image decomposition,” in IEEE International Conference on Computer Vision, 2015.