CGI2016 manuscript No.
(will be inserted by the editor)

Light Mixture Intrinsic Image Decomposition Based on a Single RGB-D Image

Guanyu Xing · Yanli Liu · Wanfa Zhang · Haibin Ling

Abstract We propose a novel intrinsic image decomposition method based on a single RGB-D image. We first separate the shading image into an illumination color component, a distant shading component and a local shading component, inducing a novel intrinsic image model that can encode color and spatial variation of scene illumination. Unlike previous methods, which assume illumination color is white, our light mixture model encodes scene illumination with two different light types, and an automatical strategy is proposed to calculate the color of the two light types. We also adopt physical based illumination prior to infer the distant shading component. To do so, we firstly recover the illumination distribution of the distant light sources through solving a system of linear equations with sparse and non-negative constraints. Then, the recovered illumination is used to synthesize a coarse distant shading image jointly with the depth map. Laterly, the synthetic image is employed as an additional constraint of distant shading component. To reduce noise disturbance from the synthetic distant shading image, a novel sampling strategy was proposed. Finally, we consider the similarity of material locally and globally, which gives reliable constrains to the reflectance component. Experimental results demonstrate the validity and flexibility of our approach.

Guanyu Xing · Wanfa Zhang
School of Computer Science & Engineering, University of Electronic Science & Technology of China (UESTC), Chengdu, China
Center for Robotics, UESTC, Chengdu, China
E-mail: xingguanyu@uestc.edu.cn

Guanyu Xing · Haibin Ling
Center for Data Analytics and Biomedical Informatics, Dept. of Computer & Information Sciences, Temple University, Philadelphia, PA, the United States

Yanli Liu
College of Computer Science, Sichuan University, Chengdu, China

Keywords intrinsic image · single RGB-D image · light mixture · physical based illumination prior

1 Introduction

Intrinsic image decomposition addresses the problem of separating a photo into the product of an illumination component that represents lighting effects and a reflectance component that is the color of the observed material. The decomposition results are of importance in many computer graphic and computer vision tasks, such as segmentation, re-lighting and color constancy.

Since for each pixel, the number of the unknowns is twice the number of measurements, intrinsic image decomposition is essentially a highly ill-posed problem. A nature and extensively adopted way is to incorporate some priors including assumptions or constraints into illumination component, reflectance component or both of them. Among various priors, the Retinex model [13] which assumes that small gradients correspond to illumination and large gradients correspond to reflectance is a classical and widely used assumption. However, although the Retinex works well in a Mondrian (i.e. piecewise constant) world, it is known to break down in the presence of occlusions, shadows, and other phenomena commonly encountered in real-world scenes [10]. More recently, researchers demonstrated that the basic Retinex model can be improved by adopting non-local texture constraints [20] or global sparsity priors [9, 17]. Bell et al. proposed a dense CRF (conditional random field) formulation based method that achieves better decomposition than previous methods on the designed database which contains a large amount of indoor images (Fig. 1(b)) [6]. Alternatively, Sinha et al. searched for global consistency based on local gray-level junction analysis to classify edges as illumination or reflectance, which can help recovering reflectance

(a) input (b) Bell et al. [6] (c) Barron et al. [3] (d) Chen et al. [8] (e) our method

Fig. 1: A test scene from the NYU dataset [18]. (a) Input color and depth image, white pixels in depth image are incomplete areas whose depth values are not recorded. ((b)∼(d)) reflectance and shading images estimated by three recent approaches for intrinsic decomposition of RGB-D images and by our method. Our method can produce nearly uniform shading for points with similar lighting condition but different material (the quilt and white fabric on the bed), and prevent the interference from the noise of the depth image.

and illumination in a world of painted polyhedra [19]. To process real images, machine learning has also been used to classify each image derivative [4,21]. Although these works have made great progress in intrinsic image decomposition, high quality decomposition for natural complex scenes is still very challenging, in which the main difficulties include (not limited to): firstly, real scenes are usually illuminated by colored light sources, some of them even illuminated by multiple colored light sources. None of the methods mentioned above considers this case. For colored image, they usually process different channels separately, which is hard to accurately recover shading images for scenes illuminated by mixed color lights. Therefore, most shading results demonstrated in previous papers are gray-scale images. Secondly, the lack of enough scene information makes the above methods fail to accurately decompose images of scenes with complex material, geometry or illumination condition.

Recently, facilitated by the development of commercial depth sensors such as Kinect [12], the depth information of scenes can be easily acquired and has been introduced into intrinsic image decomposition. For example, Lee et al. [14] tried to estimate intrinsic images from an RGB-D video captured by a moving camera based on the Retinex model. In the approach, the depth map is used to enforce non-local constraint among different surface points in the shading components. However, the approach still ignores colored light sources, and performs better for RGB-D videos instead of RGB-D images. Another practical and important problem of depth maps is: they are usually noisy and incomplete due to sensor noise, dark objects, occlusion of the structured lights, and so on, making the geometry information unreliable. Therefore, more intensive study on how to use depth map to decompose accurate intrinsic images is required.

The work most related to ours includes the methods proposed in [3,8], which also decompose intrinsic images from a single RGB-D image. In [3], Barron et al. presented the scene-SIRF model to obtain an optimized depth image, a reflectance image and a spatially varying model of illumination from a single RGB-D image. However, it needs to solve a very complex non-convex optimization problem to get the intrinsic components, making the method time-consuming. Moreover, the method performs poorly for points with inaccurate depth values, i.e., points whose depth values are missing or position outside the depth sensor's range. As a result, the shading image recovered by the method may contain unfaithful areas (The wall lamps are illegible in the shading image of Fig. 1 (c)). Chen's method [8] considers both direct and indirect illumination of a shading image, the color of illumination is also modeled in their method. However, it does not consider the mixture of different lights and only exploits the smoothness of illumination color as a constraint, which is not reliable enough. The non-local shading constraint based on the depth information is also hard to guarantee surface points with accurate shading value, due to the fact that shading of objects is decided jointly by geometry and illumination (The white fabric on bed are too bright comparing to its surrounding area in the shading image of Fig. 1 (d)).

In this paper, we present a novel framework of intrinsic image decomposition from a single RGB-D image automatically. Unlike previous work, our intrinsic image model encodes the multiple light sources with different colors. To account for the light mixture and spatial variation of illumination condition, we first deduce a new intrinsic image model which decomposes the shading into three constituent components, i.e., an illumination color component, a distant shading component and a local shading component. Then we present an iterative strategy to estimate the illumination color automatically. To estimate the illumination distribution of distant light sources, the sparse and nonnegative

constraints of illumination distribution are utilized. Then a coarse shading image is synthesized jointly by recovered illumination and depth map and further set as an additional constraint when calculating the distant irradiance component. A sampling strategy is proposed to reduce the noise disturbance of the input depth map. For the reflectance image, we introduce both the local and global material similarity to give reliable constrains to the reflectance image.

The main contributions of the paper include: (1) a new light mixture intrinsic image model for scenes with two light types, enabling the method competent for practical complex scenes; (2) a novel framework to infer intrinsic images with physical based illumination prior; and (3) an effective strategy to prevent the final shading image from noise disturbance of an input depth map. These contributions lead to an efficient and accurate approach for intrinsic images decomposition based on a single RGB-D image.

2 Model

The intrinsic image estimation is try to decompose the image into the product of a reflectance image A and a shading image. For an RGB image I we have:

$$\mathbf{I}_p(c) = \mathbf{S}_p(c)\mathbf{A}_p(c), for\ c \in \{R, G, B\} \tag{1}$$

where $\mathbf{I}_p$ is a 3×1 vector containing the observed RGB color of pixel p, $\mathbf{S}_p$ and $\mathbf{A}_p$ are also 3×1 vectors containing shading and reflectance at pixel p respectively.

Our light mixture model assumes two light types in a scene, which is consistent with practical scene configurations involving indoor/outdoor or flash/ambient lighting [11], therefore:

$$\mathbf{S}_p(c) = S_{<1,p>}\mathbf{L}_1(c) + S_{<2,p>}\mathbf{L}_2(c)$$

where $\mathbf{L}_1$ and $\mathbf{L}_2$ are 3×1 vectors representing the illumination color, $S_{<1,p>}$ and $S_{<2,p>}$ are the values of p at the two gray-scale shading images corresponding to different light types, then we have:

$$\mathbf{I}_p(c) = (S_{<1,p>}\mathbf{L}_1(c) + S_{<2,p>}\mathbf{L}_2(c))\mathbf{A}_p(c) \tag{2}$$

Many intrinsic image decomposition methods such as [2, 14] assume that the light sources are distant from the examined scene; consequently, points with the same normal directions will share uniform shading value if unoccluded. As there may be some local light sources and occlusions in the real world, the scene's illumination may vary spatially in practice. Obviously, typical decomposition models cannot handle these cases. Our approach considers all factors described above. We use a distant shading image D to encode the scene illuminated only by distant light sources and ignore occlusions, then Eq. 2 can be written as:

$$\mathbf{I}_p(c) = (\frac{S_{<1,p>}}{D_p}\mathbf{L}_1(c) + \frac{S_{<2,p>}}{D_p}\mathbf{L}_2(c))D_p\mathbf{A}_p(c)$$

denote $\frac{S_{<1,p>}}{D_p}$, $\frac{S_{<2,p>}}{D_p}$ as $k_{<1,p>}$ and $k_{<2,p>}$ respectively,

$$\mathbf{I}_p(c) = (k_{<1,p>}\mathbf{L}_1(c) + k_{<2,p>}\mathbf{L}_2(c))D_p\mathbf{A}_p(c) \tag{3}$$

Let $\mathbf{W}_p(c) = \frac{k_{<1,p>}+k_{<2,p>}}{k_{<1,p>}\mathbf{L}_1(c)+k_{<2,p>}\mathbf{L}_2(c)}$ ($\mathbf{W}_p$ is a 3×1 vector), multiply $\mathbf{W}_p(c)$ in both sides of Eq. 3, we have $\mathbf{W}_p(c)\mathbf{I}_p(c) = (k_{<1,p>} + k_{<2,p>})D_p\mathbf{A}_p(c)$, therefore,

$$\mathbf{I}_p(c) = (k_{<1,p>} + k_{<2,p>})D_p\frac{1}{\mathbf{W}_p(c)}\mathbf{A}_p(c) \tag{4}$$

Denote $(k_{<1,p>} + k_{<2,p>})$ and $\frac{1}{\mathbf{W}_p(c)}$ as K_p and $\mathbf{C}_p(c)$ respectively, our final model is

$$\mathbf{I}_p(c) = D_pK_p\mathbf{C}_p(c)\mathbf{A}_p(c), for\ c \in \{R, G, B\} \tag{5}$$

From the definition of the four items in the model, C encodes scene's illumination color, we call it as the color image; K encodes the spatial variation of illumination caused by light sources near the examined scene and occlusions, which is called as the local shading image; D and A are the distant shading image and the reflectance image respectively. We will discuss how to calculate them in the rest of this article.

3 Distant illumination recovery

For intrinsic image decomposition, a more reliable constraint of shading can be established if we have known scene illumination, due to the truth that shading value of a point in a scene is decided jointly by geometry and lighting condition. Therefore, this section will discuss the method of distant illumination recovery.

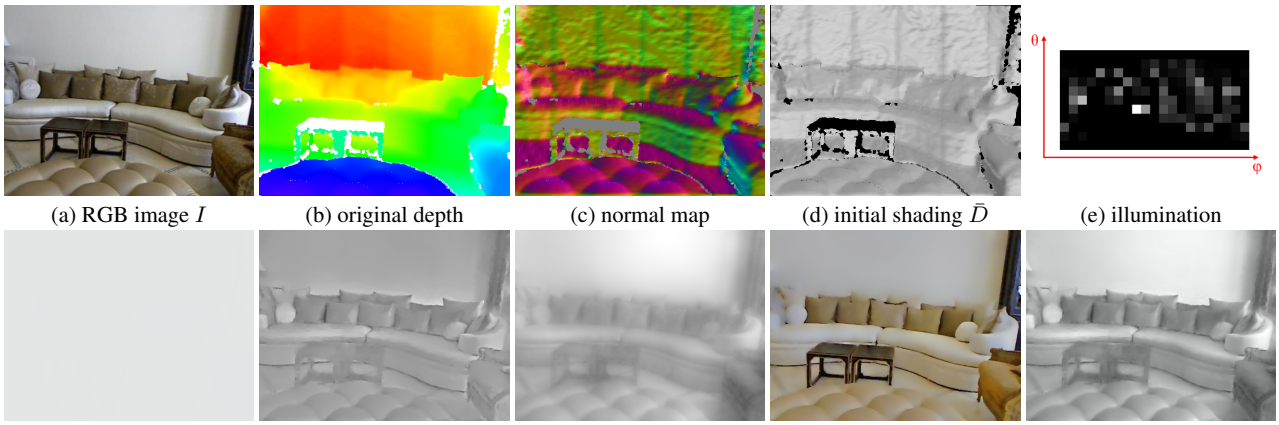
The color of illumination is described as a separate component in our intrinsic image model, so we only consider light sources with white color in this section. The assumption of white illumination indicates that recovering the intensity of every light source is the primary goal of our lighting estimation algorithm.

Denote the intensity and direction vector of a distant light source $L^{\mathbf{d}}$ as L and $\mathbf{d}$ respectively. Then we can produce a shadow removed image $I^{\mathbf{d}}$ of the examined scene corresponding to $L^{\mathbf{d}}$ as following:

$$\mathbf{I}^{\mathbf{d}}_p(c) = max(0, L\mathbf{A}_p(c) < \mathbf{d}, \mathbf{n}_p >), for\ c \in \{R, G, B\},$$

where $\mathbf{n}_p$ is the reflectance and normal vector of point p, $<>$ is the dot product of two vectors. For convenience, we call $I^{\mathbf{d}}/L$ basis image, which is related to the light source $L^{\mathbf{d}}$. Obviously, an image can be written as a linear combination of basis images corresponding to different distant light sources under the assumption that light sources are far away from the examined scene. As an approximation, we sample several directions of light sources, therefore:

$$\mathbf{I}_p(c) = \sum_{q \in \aleph_d} L_q\mathbf{A}_p(c) \cdot max(0, < \mathbf{d}_q, \mathbf{n}_p >), \tag{6}$$



(a) RGB image I (b) original depth (c) normal map (d) initial shading $\bar{D}$ (e) illumination

(f) light color C (g) distant shading D (h) local shading K (i) reflectance A (j) shading S

Fig. 2: The input RGB-D image is demonstrated in (a) and (b), white pixels in (b) are incomplete areas whose depth values are not recorded. (c) shows the normal map recovered from the depth map. (d) is the initial distant shading image $\bar{D}$ generated from the recovered distant illumination distribution which is presented in (e). (f)∼(i) are the recovered illumination color image, distant shading image D, local shading image K and reflectance image A respectively. (j) is the final shading image S which is the product of C, K and D.

where $\aleph_d$ is the set of sampling light sources. We sample the direction of light source as all combinations of polar angles θ from $[-\frac{\pi}{2}, \frac{\pi}{2}]$ and azimuth angles φ from $[0, 2\pi]$, at intervals of $\pi/10$. Divided by $\mathbf{A}_p(c)$ in both sides, Eq. 6 becomes

$$\sum_{q \in \aleph_d} L_q \cdot max(0, < \mathbf{d}_q, \mathbf{n}_p >) - \frac{\mathbf{I}_p(c)}{\mathbf{A}_p(c)} = 0, \tag{7}$$

Note that this is a linear equation with unknowns of L_q and $\frac{1}{\mathbf{A}_p(c)}$, if we choose n sample points in the scene, a system of linear equations can be constructed. However, solving the linear equations may get zero solutions, to avoid this, we let $\frac{1}{\mathbf{A}_p(c)} = 1 + \rho_p$, where ρ_p is a non-negative value, for the reason that $\mathbf{A}_p$ is the reflectance parameter which is less than 1. Then, we have

$$\sum_{q \in \aleph_d} L_q \cdot max(0, < \mathbf{d}_q, \mathbf{n}_p >) - \mathbf{I}_p(c)\rho_p = \mathbf{I}_p(c), \tag{8}$$

Unfortunately, this system of linear equations is still undeterministic, because there are n equations with $n + |\aleph_d|$ unknowns, which is under constrains. Fortunately, different points in the scene may share the same material, this will reduce the number of unknowns dramatically. To find points with similar material, we cluster the input image by mean shift in the chromaticity domain, and assume points belonged to same cluster share similar material [8, 14].

Additional constraints are also employed to improve the accuracy of the approximation. A non-negative constraint is enforced. That is, $L_q \geq 0$ for $q \in \aleph_d$ and $\rho_p > 0$. Another constraint is the sparse representation saying that images produced by a Lambertian scene can be efficiently represented by a sparse set of images generated by directional light sources [16].

Theoretically, we can select any channel of the image to create the equations. We found by experiment that using the shading image recovered by Retinex model [13] as the input image I yields a better performance than other models. Fig. 2 (e) demonstrates the estimated illumination distribution, from which we can see the estimated lighting parameters are very sparse.

With the estimated illumination, we can produce an initial distant irradiance image $\bar{D}$. For a pixel p, we have:

$$\bar{D}_p = \sum_{q \in \aleph_d} L_q \cdot max(0, < \mathbf{d}_q, \mathbf{n}_p >), \tag{9}$$

In fact, $\bar{D}$ will be the distant shading image if using the precise normal map, unfortunately, the depth map captured by current commercial sensor is noisy and incomplete, which makes $\bar{D}$ imperfect. Fig. 2 (d) demonstrates the shading image produced by the original depth map. Our method try to combine the Retinex model and the synthetic shading image $\bar{D}$ to produce a better decomposition result.

4 Intrinsic components recovery

Our intrinsic image decomposition method contains two steps. The first step is to estimate the color image C; the second step solves K, D and A.

4.1 Color image estimation

We first discuss how to estimate the color image C. Adopting all symbols defined in Sec. 2, for a pixel p in channel $c \in \{R, G, B\}$, $\mathbf{C}_p(c) = \frac{k_{<1,p>}\mathbf{L}_1(c) + k_{<2,p>}\mathbf{L}_2(c)}{k_{<1,p>} + k_{<2,p>}}$, denote $\frac{k_{<1,p>}}{k_{<1,p>} + k_{<2,p>}}$ as α_p which is called as light mixture parameter, we have:

$$\mathbf{C}_p = \alpha_p \mathbf{L}_1 + (1 - \alpha_p)\mathbf{L}_2, \tag{10}$$

The estimation of color image is converted to a problem of solving α_p and illumination color $\mathbf{L}_1$, $\mathbf{L}_2$. The similar problem has been discussed in Hsu's paper, however, the illumination color is indicated by user [11]. We seek an automatic way in this paper.

4.1.1 Illumination color calculation

We use a simple method to give a coarse estimation of $\mathbf{L}_1$ and $\mathbf{L}_2$, then refine them by an iteration procedure.

Illumination color initialization. We initialize illumination color based on real life experiences. We set $\mathbf{L}_1 = (1,1,1)^T$, due to the fact that white light sources are very common in our life. We also notice that the specular reflection has less effect in changing the color of incident light than that of diffused reflection, therefore we adopt pixels belonging to highlight areas to calculate $\mathbf{L}_2$, we first get the principal components of all highlight pixels' RGB value vector by PCA (Principal component analysis), then set $\mathbf{L}_2$ as the first principal component. The highlight pixels are detected by Zhai's method [22]

Illumination color refinement. We use a two step iterative strategy to refine the initial illumination color.

Step1: We adopt the method described in [11] to estimate material colors of the scene, a voting scheme is proposed to decide the final material of pixels in the image. This method cannot distinguish albedos differing only by a scale factor, i.e. two albedo values (a,b,c) and (ka,kb,kc) will be treated as the same. Therefore, Eq. 2 can be wrote as:

$$\mathbf{I}_p(c) = (\frac{S_{<1,p>}}{\tilde{k}_p}\mathbf{L}_1(c) + \frac{S_{<2,p>}}{\tilde{k}_p}\mathbf{L}_2(c))(\tilde{k}_p\mathbf{A}_p(c)),$$

where $c \in \{R,G,B\}$, $\tilde{k}_p$ is the factor to make two albedo values differed by a scale have the same value. Let $\tilde{\mathbf{A}}_p = \tilde{k}_p\mathbf{A}_p$, $k^{'}_{<1,p>} = \frac{S_{<1,p>}}{\tilde{k}_p}$ and $k^{'}_{<2,p>} = \frac{S_{<2,p>}}{\tilde{k}_p}$, we have:

$$\mathbf{I}_p(c) = (k^{'}_{<1,p>}\mathbf{L}_1(c) + k^{'}_{<2,p>}\mathbf{L}_2(c))\tilde{\mathbf{A}}_p(c) \tag{11}$$

Obviously, $k^{'}_{<1,p>}$ and $k^{'}_{<2,p>}$ can be also calculated if known illumination and material color. In addition, this method cannot get material color of all pixels in the image, but it will not influence the calculating of illumination color.

Step2: We sample n pixels randomly from the image, for the ith pixel, we set $(\frac{\mathbf{I}_{p_i}(R)}{\tilde{\mathbf{A}}_{p_i}(R)}, \frac{\mathbf{I}_{p_i}(G)}{\tilde{\mathbf{A}}_{p_i}(G)}, \frac{\mathbf{I}_{p_i}(B)}{\tilde{\mathbf{A}}_{p_i}(B)})$ as the ith row of an $n \times 3$ matrix $\tilde{\mathbf{I}}$. According to Eq. 11, $\tilde{\mathbf{I}} = \tilde{\mathbf{K}} \cdot \tilde{\mathbf{L}}$, where $\tilde{\mathbf{K}}$ is an $n \times 2$ matrix with $(k^{'}_{<1,p_i>}, k^{'}_{<2,p_i>})$ as its ith row, $\tilde{\mathbf{L}}$ is a 2×3 matrix which records the illumination colors. The two matrix $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{L}}$ can be solved by NMF (non-negative matrix factorization) [5], the results from step1 are used as the initialization of the NMF procedure.

4.1.2 Calculation of light mixture parameter

We follow Hsu's method [11] in this section, and get the expression below:

$$\mathbf{I}^{'}_p = \alpha_p\mathbf{A}^{'}_p * \mathbf{L}^{'}_1 + (1-\alpha_p)\mathbf{A}^{'}_p * \mathbf{L}^{'}_2, \tag{12}$$

where α_p is the light mixture parameter, $\mathbf{I}^{'}_p = [\frac{\mathbf{I}_p(R)}{\mathbf{I}_p(B)}, \frac{\mathbf{I}_p(G)}{\mathbf{I}_p(B)}]^{'}$, $\mathbf{A}^{'}_p = [\frac{\mathbf{A}_p(R)}{\mathbf{A}_p(B)}, \frac{\mathbf{A}_p(G)}{\mathbf{A}_p(B)}]^{'}$ and $\mathbf{L}^{'}_i = [\frac{\mathbf{L}_i(R)}{\mathbf{L}_i(B)}, \frac{\mathbf{L}_i(G)}{\mathbf{L}_i(B)}]^{'}$ $(i \in \{1,2\})$, the symbol $*$ is the Hadamard product.

Note that the solving of α_p is similar to the classical foreground/backgroud mixture problem in matting, therefore, we adopt the matting algorithm proposed by Levin et. al [15] to solve this problem.

4.2 Shading and reflectance components estimation

Divide the estimated color image in both sides of Eq. 5. For convenience, we still denote $\mathbf{I}_p(c)/\mathbf{C}_p(c)$ as $\mathbf{I}_p(c)$ for $c \in \{R,G,B\}$, therefore:

$$\mathbf{I}_p(c) = K_pD_p\mathbf{A}_p(c), \tag{13}$$

As is common in intrinsic image decomposition, we conduct decomposition in the logarithmic domain. Taking logarithms on both sides yields:

$$\mathbf{i}_p(c) = k_p + d_p + \mathbf{a}_p(c), \tag{14}$$

Then we formulate the decomposition as a problem of minimizing the following energy function:

$$E(k,d,\mathbf{a}) = E_{data}(k,d,\mathbf{a}) + E_A(\mathbf{a}) + E_D(d) + E_K(k), \tag{15}$$

We call E_{data} as data term, E_A as reflectance term, E_D the distant shading term, while E_K the local shading term. They are described in detail in the next few sections.

4.2.1 Data term

This term is to make sure that the original image I can be reconstructed by the recovered components. Our data term is defined as:

$$E_{data}(k,d,\mathbf{a}) = \sum_{c\in\{R,G,B\}}\sum_{p\in\aleph_I}(i_p - k_p - d_p - \mathbf{a}_p(c))^2, \tag{16}$$

where $\aleph_I$ is the set of all pixels in image I.

4.2.2 Reflectance term

We consider both local and global similarity of material in our method. The reflectance term is defined as:

$$E_A(\mathbf{a}) = \lambda_A^l E_A^l(\mathbf{a}) + \lambda_A^g E_A^g(\mathbf{a}), \tag{17}$$

where λ_A^l and λ_A^g are weight parameter, $E_A^l(\mathbf{a})$ and $E_A^g(\mathbf{a})$ are local constraint term and global constraint term respectively.

Local constraint term. This term comprises pairwise terms that penalize differences between adjacent pixels in I:

$$E_A^l(\mathbf{a}) = \sum_{c\in\{R,G,B\}} \sum_{p\in\aleph_I} \sum_{q\in N_p} \alpha_{p,q}^l (\mathbf{a}_p(c) - \mathbf{a}_q(c))^2, \tag{18}$$

where $\aleph_I$ is the set of all pixels in image I, N_p is the 5×5 neighborhood of pixel p, $\alpha_{p,q}^l$ is calculated by:

$$\alpha_{p,q}^l = e^{-1\cdot\kappa_1\cdot\|ch(I_p)-ch(I_q)\|} min(1, \sqrt{\kappa_2 \cdot lum(I_p) lum(I_q)}), \tag{19}$$

where $ch(I_p)$ and $lum(I_p)$ denote the chromaticity and luminance of p, κ_1, κ_2 are two parameters, the luminance and chromaticity values of pixels are calculated in the HSV color space. The left term in this equation expresses the truth that pixels with similar chromaticity value are likely to have similar material. κ_1 can adjust the sensitivity to pixels' chromaticity variation of our method, we set it as 200 for most cases in this paper, a larger value should be adopted for scenes with complex texture. The right term is used to penalize pixels with low luminance value, due to the fact that much noise exists in dark areas. κ_2 can control the strength of the punishment, we set it as 1 for photos of real scenes, while 100 for virtual scenes, because there is little noise in rendered images.

Global constraint term. This term tries to keep non-adjacent pixels but with similar material share the same value in the estimated reflectance image. To reduce the dimensionality of the problem, we only select one matched pixel for each pixel in the image. The material similarity of two non-neighbored areas can be maintained by the local constraint by setting the weight parameter λ_A^g a very big value. Next we will describe the strategy of matched pixel selection.

In Sec. 4.1.1, we have recovered material color of part pixels in the image, for these pixels, we select pixel with similar material color and chromaticity value but farthest distance in the image plane as the matched pixel $p^{'}$; for other pixels, we only use small difference in chromaticity value and the farthest distance as the judgements. The definition of the global constraint term $E_A^g(\mathbf{a})$ is:

$$E_A^g(\mathbf{a}) = \sum_{c\in\{R,G,B\}} \sum_{p\in\aleph_I} \alpha_{p,p'}^g (\mathbf{a}_p(c) - \mathbf{a}_{p'}(c))^2, \tag{20}$$

where $\alpha_{p,p'}^g = e^{-20\|\mathbf{I}_p - \mathbf{I}_{p'}\|}$. This weight parameter is used to make sure that pixel pairs with similar RGB, chromaticity and material color contribute more to the global material similarity.

4.2.3 Distant shading term

The distant shading term comprises two components: one for smoothness, the other for keeping the recovered shading image close to the synthetic shading $\bar{D}$. Denote them as smoothness term E_D^s and initial constraint term E_D^i respectively.

$$E_D(d) = \lambda_D^s E_D^s(d) + \lambda_D^i E_D^i(d), \tag{21}$$

where λ_D^s and λ_D^i are weight parameters. We now describe these two terms.

Smooth term. Note that our distant shading image D ignores occlusions and highlight in the scene, so if two adjacent points have similar normals, we expect them to have similar shading. The smooth term is designed to model the angular coherence of distant illumination, it has the following form:

$$E_D^s(d) = \sum_{p\in\aleph_I} \sum_{q\in N_p} \beta_{p,q}^s (d_p - d_q)^2, \tag{22}$$

The definition of $\aleph_I$ and N_p is similar as them in Eq. 18. The weight $\beta_{p,q}^s$ is defined as:

$$\beta_{p,q}^s = e^{(-100\|\mathbf{n}_p - \mathbf{n}_q\|)}, \tag{23}$$

This weight function is used to penalize pixels with different normals.

Initial constraint term. In Sec. 3, we have generated an initial distant shading image $\overline{D}$ which can be used as a constraint term in the energy function. To minimize the disturbance from $\overline{D}$, we just let the unknown shading image close to the initial image $\overline{D}$ in several sampling pixels, the value of rest pixels can be estimated according to the smoothness of shading and reflectance. The initial constraint term is defined as:

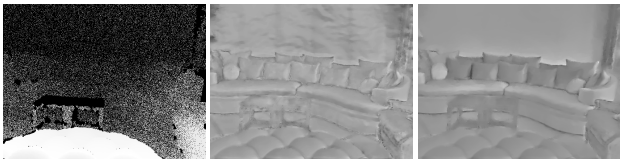
$$E_D^i(d) = \sum_{p\in\aleph_{init}^d} \beta_p^i (d_p - \bar{d}_p)^2, \tag{24}$$

where $\bar{d}_p$ is the logarithm of $\bar{D_p}$, $\aleph_{init}^d$ is the set of sampling pixels. The value of the weighting parameter β_p^i is calculated as:

$$\beta_p^i = \begin{cases} \frac{dep_p}{dis_{min}} & dep_p < dis_{min} \\ 0.8 + 0.2\frac{dis_{max} - dep_p}{dis_{max} - dis_{min}} & dis_{min} \le dep_p \le dis_{max} \\ 0.8 - 0.8\frac{dep_p - dis_{max}}{dis_{max}} & dis_{max} \le dep_p \le 2dis_{max} \\ 0 & dep_p \ge 2dis_{max} \end{cases} \tag{25}$$

where dep_p is the depth value of p, dis_{min} and dis_{max} together decide the effective range of the depth sensor. Eq. 25 penalize points outside the effective range, because their depth value are very inaccurate. Our approach sets dis_{min}, dis_{max} as 1.2m and 4.0m, which correspond to the range of Kinect sensor.

To get the pixel set $\aleph^d_{init}$, we propose a sampling strategy. We first remove pixels belonging to incomplete areas of the depth image, then using the weight parameter β^i_p calculated by Eq. 25 to decide whether a pixel should be sampled or not. For a pixel p, let $\eta_p = e^{-5(1-\beta^i_p)}$, and generate a random variable σ_p in the range of $[0, 0.7]$ according to the uniform distribution, p will be add into $\aleph^d_{init}$, if $\eta_p > \sigma_p$. The sampling result is demonstrated in Fig. 3 (a). It shows that points with accurate depth value have higher sampling rate, while lower rate for points far away from the camera. Fig. 3 demonstrates the calculated distant shading images with (Fig. 3. (b)) and without (Fig. 3. (c)) using the proposed sampling strategy, which proves that our sampling strategy can reduce noise of the captured depth map effectively.



(a) sampling result (b) without sampling (c) with sampling

Fig. 3: Comparison between the calculated distant shading images of the scene demonstrated in Fig. 1 with and without using our sampling strategy. (a) exhibits the result of our sampling strategy. (b) and (c) demonstrate the calculated distant shading images with and without using the proposed sampling strategy.

4.2.4 Local shading term

Similar as the distant shading term, our local shading term contains two components: a smooth component and an initial constraint component. Therefore,

$$E_K(k) = \lambda^s_K E^s_K(k) + \lambda^i_K E^i_K(k), \tag{26}$$

Smooth term. This term tries to keep the coherence of local illumination, adopting the symbols used in previous sections, $E^s_K(k)$ is defined as:

$$E^s_K(k) = \sum_{p\in\aleph_I} \sum_{q\in N_p} (k_p - k_q)^2, \tag{27}$$

Initial constraint term. This term helps to capture the occlusion and highlight in the scene. First, we need to get an initial local shading image $\bar{K}$. According to Eq. 11, $k'_{<1,p>} + k'_{<2,p>} = (S_{<1,p>} + S_{<2,p>})/\tilde{k}_p$, note that $S_{<1,p>} + S_{<2,p>}$ is p's final shading value. We make an approximation and let our initial local shading image $\bar{K}$ be:

$$\bar{K}_p = \begin{cases} \frac{k'_{<1,p>} + k'_{<2,p>}}{\bar{D}_p} & p \in \aleph^l_{init} \\ 0 & otherwise \end{cases} \tag{28}$$

where $\bar{D}$ is the initial distant shading image described in Sec. 3, $\aleph^l_{init}$ is the set of pixels whose material color can be recovered and depth value is recorded by the depth sensor.

The initial constraint term is defined as:

$$E^i_K(k) = \sum_{p\in\aleph^l_{init}} (k_p - \bar{k}_p)^2, \tag{29}$$

where $\bar{k}_p = ln(\bar{K}_p)$. Since $\bar{K}$ is just an approximation of the real local shading image, we set λ^i_K to a small value comparing to other weight parameters.

5 Experiments

We use the MPI-Sintel dataset [7] to evaluate our intrinsic image decomposition algorithm, the weight parameters are set as $\lambda^l_A = 15$, $\lambda^g_A = 200$, $\lambda^s_D = 1$, $\lambda^i_D = 1$, $\lambda^s_K = 0.5$, $\lambda^i_K = 0.01$ in our experiments. We select 8 different scenes, and obtain reflectance and shading images using our approach with the whole pipeline, our approach but without considering illumination color (denoted as NC), the approach of Chen et al. [8], the approach of Barron et al. [3] and the approach of Bell et al. [6]. For comparison, we make a quantitative evaluation of the results obtained by the different approaches. The calculated LMSE (Local mean square error) [10] of these methods are demonstrated in Tab. 1, from which we can see that our approach performs better than other methods, and the adopting of illumination color makes the algorithm more accurate. Note that the LMSE of the bamboo scene is smaller when we do not consider illumination color, it is mainly caused by the bamboo, which makes our recovered illumination color a little green. Fig. 4 demonstrates the comparison results between the recovered intrinsic images and the ground truth of two test scenes (left is the market scene and right is the bamboo scene), we find that the intrinsic components produced by our method are very close to the ground truth, although the colors of intrinsic images of the bamboo scene are a little different from the ground truth, there are inter-reflections between the green bamboos, our results are still reasonable.

We also test our algorithm on the NYU-date set which contains RGB-D images captured by Kinect. Original raw sensor depth maps without any repairing are adopted in all of our experiments. The estimated intrinsic components are compared with the intrinsic images produced by Bell's [6], Chen's [8] and Barron's [3] methods. The comparison results are demonstrated in Fig. 1 and Fig. 5, from which we

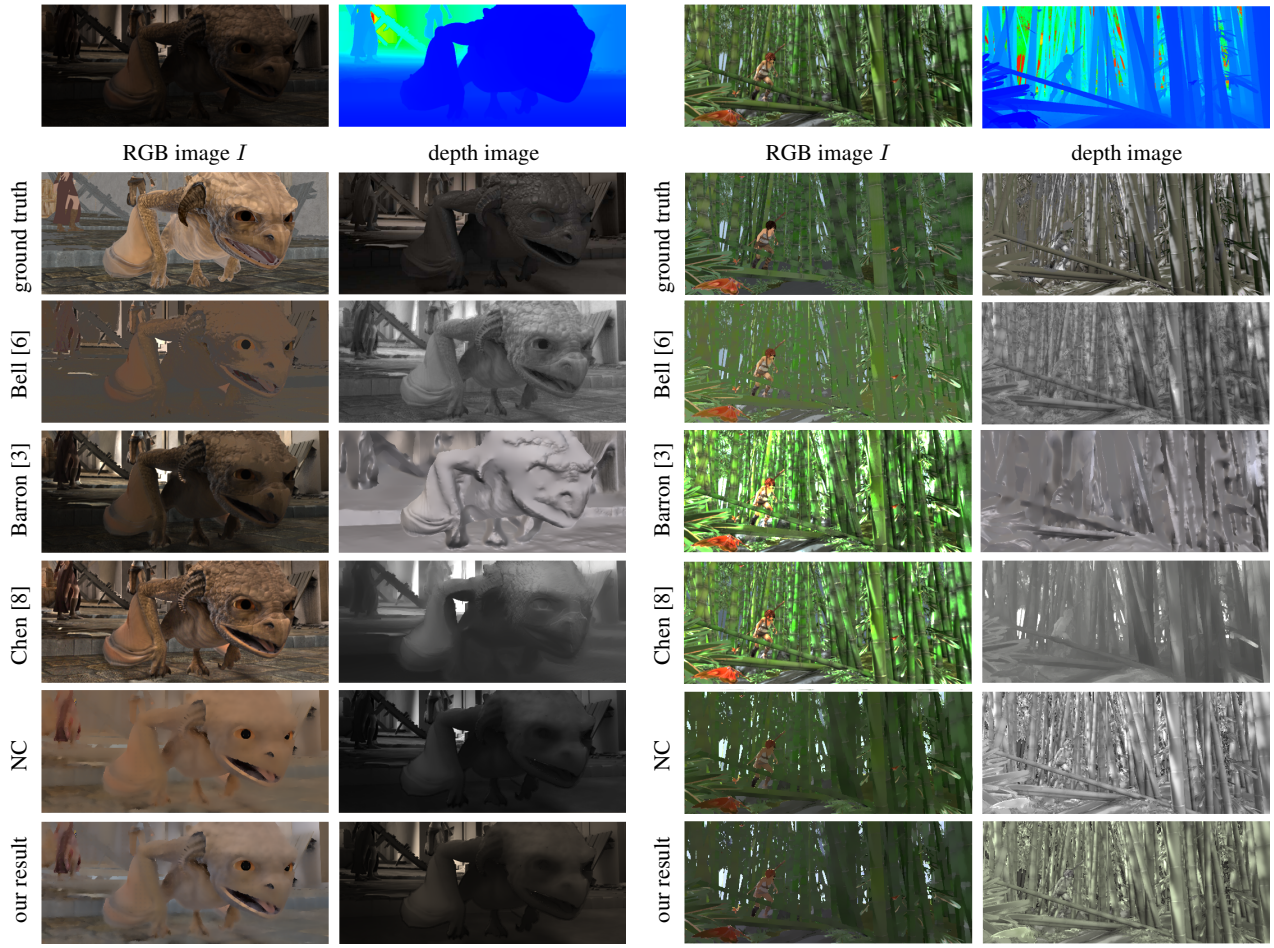


Fig. 4: Comparison between intrinsic images estimated by different methods and the ground truth.

Scene	Bell [6]	Barron [3]	Chen [8]	NC	Ours
allry1	0.0597	0.0550	0.0615	0.0399	**0.0397**
allry2	0.0459	0.0506	0.0435	0.0317	**0.0242**
ambush	0.1470	**0.1147**	0.1439	0.1664	0.1423
bamboo	**0.0480**	0.0869	0.0835	0.0669	0.0692
bandage	0.0762	0.0804	0.0783	0.0744	**0.0727**
market	0.0572	0.1254	0.0795	0.0403	**0.0381**
shaman	0.1122	0.1554	0.1420	0.1058	**0.1019**
sleeping	0.0300	0.0438	0.0323	0.0348	**0.0248**

Table 1: Quantitative evaluation of the albedo and shading images estimated by different approaches on the MPI-Sintel dataset. We adopt LMSE as our error metric.

can see the reflectance images recovered by Bell's method are too smooth to make the shading image accurate; Chen's method makes objects with small RGB value too dark while objects with large RGB value too bright in the recovered shading image; Barron's method produces unfaithful areas in the decomposed images. For example, the ornaments on the wall in Fig. 1 are hard to be recognized according to the shading image. Our method can produce nearly uniform reflectance for points with similar albedo, and the shading image still conforms with common sense. The color image can capture the light mixture of scene's illumination, take the first scene in Fig. 5 for example, the light color of the area out of the restroom is white, while the restroom is a little blue, which is consistent with our recovered color image. Decomposition results of three additional scenes are presented in Fig. 6, we can find that our color image of the second scene can capture both the colors of indoor light source (orange) and the light from the window (white).

Our algorithm is implemented on a PC with Core i7-4790 4.0GHz CPU and 16GB RAM. The average time cost of our method and Bell's method is nearly 4 minutes (for image with 561×427 resolution), while it takes 10 minutes for Chen's method and 40 minutes for Barron's. We only need to solve several optimization problems and the number of unknown parameters is much less than that in Chen's and Barron's methods, which makes our algorithm more efficient.

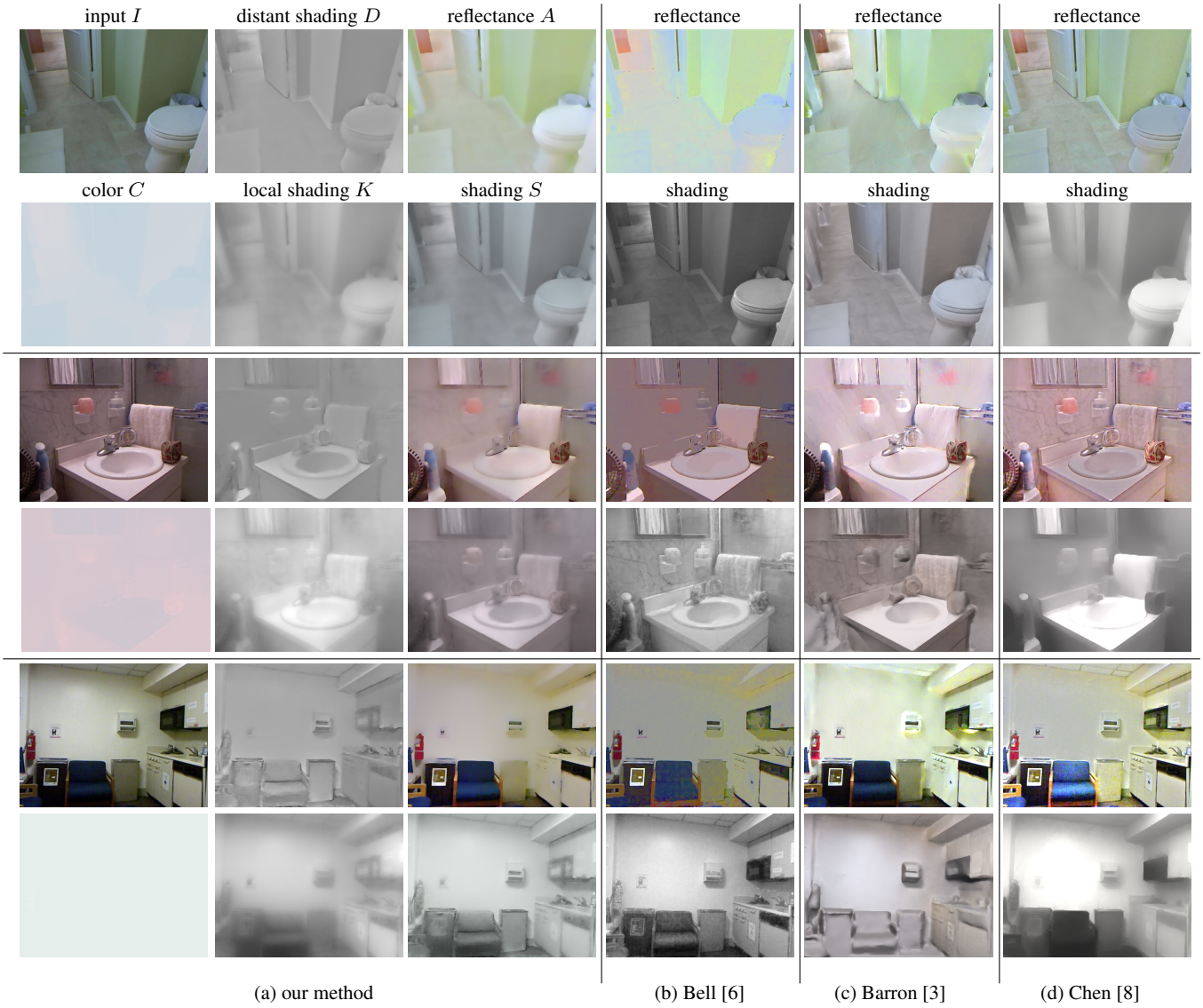


Fig. 5: Three test scenes from the NYU dataset [18]: (a) demonstrates the results of our method including: illumination color image C, the distant irradiance image D, the local irradiance image K, the reflectance image A and the shading image S. (b)∼(d) show the reflectance and shading image produced by other there intrinsic image algorithms.

6 Conclusions and future work

We have presented a new method to decompose an RGB-D image into its intrinsic components. We first propose a novel four-component intrinsic image model, which separates the shading image into illumination color component, distant shading component and local shading component. This model can describe the light mixture and spatial variation of illumination condition. A new iterative strategy is proposed to calculate the illumination color component automatically, then we estimate the other three intrinsic components through solving an energy minimization problem. Unlike previous methods, Our algorithm estimates the illumination distribution of the distant light sources through solving a system of linear equations. The recovered illumination can produce an initial distant shading image based on the depth map. To decrease the disturbance from noise of the original depth map, we adopt a sampling strategy which can ensure that pixels with accurate depth value will have higher possibility to be selected. We also employ both local and global constraints of the reflectance component, which makes the decomposed intrinsic images reliable.

Although our method runs faster than most of previous methods, the efficiency still cannot meet the practical application. Our recovered distant illumination still has error, and the accuracy of decomposition results for scenes with dark objects and complex texture can also be improved. These limitations indicate our future work.

Fig. 6: The decomposition results of three other scenes, which include: input RGB image, the color image C,the distant shading image D, the local shading image K, the reflectance image A and the shading image S.

Acknowledgements

This research is supported by National Natural Science Foundation of China (Grant No.61402081, 61572333), 863 Program of China (Grant No.2015AA016405), Fundamental Research Funds for the Central Universities (Grant No. ZYGX2014J059), China Scholarship Council (Grant No. [2015]3012) and the Oversea Academic Training Funds, UESTC. Ling was supported in part by National Science Foundation (Grant No.1449860, 1218156 and 1350521).

References

1. Barron, J.T., Malik, J.: Shape, albedo, and illumination from a single image of an unknown object. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 334–341 (2012)
2. Barron, J.T., Malik, J.: Intrinsic scene properties from a single rgb-d imag. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17–24 (2013)
3. Bell, M., Freeman, W.T.: Learning local evidence for shading and reflectance. In: Proc. of the Int. Conference on Computer Vision, pp. 670–677 (2001)
4. Bell, S., Bala, K., Snavely, N.: Learning the parts of objects by non-negative matrix factorization. Nature **401**, 788–791 (1999)
5. Bell, S., Bala, K., Snavely, N.: Intrinsic images in the wild. ACM Trans. on Graphics (SIGGRAPH) **33**(4) (2014)
6. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: A. Fitzgibbon et al. (Eds.) (ed.) European Conf. on Computer Vision (ECCV), Part IV, LNCS 7577, pp. 611–625. Springer-Verlag (2012)
7. Chen, Q., Koltun, V.: A simple model for intrinsic image decomposition with depth cues. In: Proc. of The International Conference on Computer Vision (2013)
8. Gehler, P., Rother, C., Kiefel, M., Zhang, L., Schölkopf, B.: Recovering intrinsic images with a global sparsity prior on reflectance. In: Advances in Neural Information Processing Systems (NIPS), pp. 765–773 (2011)
9. Grosse, R., Johnson, M.K., Adelson, E.H., Freeman, W.T.: Ground-truth dataset and baseline evaluations for intrinsic image algorithms. In: International Conference on Computer Vision, pp. 2335–2342 (2009)
10. Hsu, E., Mertens, T., Paris, S., Avidan, S., Durand, F.: Light mixture estimation for spatially varying white balance. In: Proc. SIGGRAPH 2008, pp. 70:1–7. Los Angeles, California, USA (2008)
11. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., Fitzgibbon, A.: Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In: Proc. UIST, pp. 559–568 (2011)
12. Land, E.H., McCann, J.J.: Lightness and retinex theory. Journal of the Optical Society of America **61**(1), 1–11 (1978)
13. 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: Proc. of The 12th European Conference on Computer Vision, pp. 327–340 (2012)
14. Levin, L., Weiss, Y.: A closed form solution to natural image matting. In: In IEEE Computer Vision and Pattern Recognition, pp. 61–68 (2006)
15. Mei, X., Ling, H., Jacobs, D.W.: Sparse representation of cast shadows via l1-regularized least squares. In: Proc. ICCV, pp. 583–590. Kyoto, Japan (2009)
16. Shen, L., Yeo, C., Hua, B.S.: Intrinsic images decomposition using a local and global sparse representation of reflectance pp. 2904–2915 (2011)
17. Silberman, N., D. Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: Proc. ECCV, pp. 746–760 (2012)
18. Sinha, P., Adelson, E.: Recovering reflectance and illumination in a world of painted polyhedra. In: Proc. of the Fourth Int. Conf. on Computer Vision, pp. 156–163 (1993)
19. Tappen, M.F., Adelson, E.H., Freeman, W.T.: Estimating intrinsic component images using non-linear regression. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1992–1999 (2006)
20. Tappen, M.F., Freeman, W.T., Adelson, E.H.: Recovering intrinsic images from a single image. IEEE Trans. Pattern Anal. Mach. Intell. **27**(9), 1459–1472 (2005)
21. Zhai, Y., Shah, M.: Visual attention detection in video sequences using spatiotemporal cues. In: Proceedings of the 14th ACM international conference on Multimedia, pp. 815–824 (2006)