Extracting Smooth and Transparent Layers from a Single Image *

Sai-Kit Yeung saikit@ust.hk
Tai-Pang Wu pang@ust.hk
Chi-Keung Tang cktang@cse.ust.hk

Vision and Graphics Group
The Hong Kong University of Science and Technology

Abstract

Layer decomposition from a single image is an under-constrained problem, because there are more unknowns than equations. This paper studies a slightly easier but very useful alternative where only the background layer has substantial image gradients and structures. We propose to solve this useful alternative by an expectation-maximization (EM) algorithm that employs the hidden markov model (HMM), which maintains spatial coherency of smooth and overlapping layers, and helps to preserve image details of the textured background layer. We demonstrate that, using a small amount of user input, various seemingly unrelated problems in computational photography can be effectively addressed by solving this alternative using our EM-HMM algorithm.

Keywords: Image decomposition, computational photography, vision for graphics.

1. Introduction

The problem of separating a set of overlapping layers from a single image is a severely under-constrained problem. Previous approaches used depth from focus [10], multiple images and motion [12], repetitive motion [9], independent component analysis [4], and sparsity priors [6]. This paper considers a slightly easier version of the problem: given a single image where only the background layer has substantial image gradients and structures, can we recover the background layer as well as the overlapping/transparent layers? That is,

$$I(x,y) = F(x,y) + \beta(x,y)B(x,y) \tag{1}$$

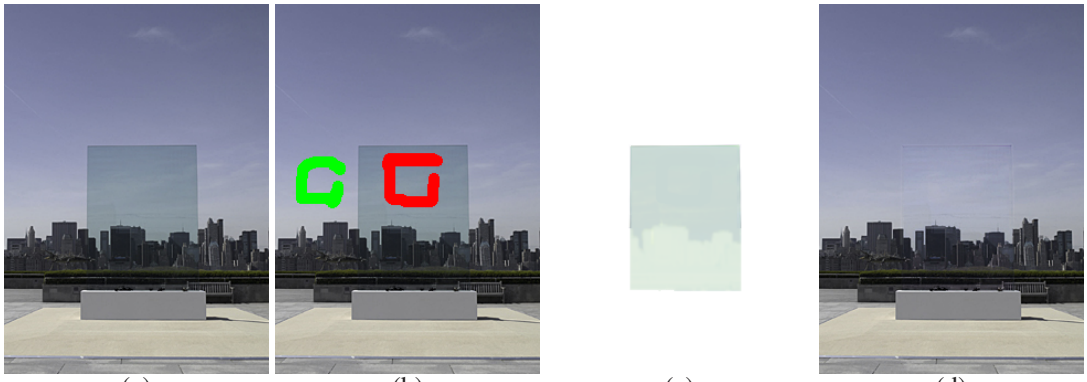
where I is the input image. F is a set of overlapping layers possibly with soft and transparent boundaries. B is the background layer, which can be attenuated by a smooth transparent layer β. I, F, B and β are RGB vectors (β is modeled to respond differently each color channel). This slightly easier problem is still ill-posed, as given a single image I there is still an infinite number of F, smooth β, and B that gives the same I. Suppose we first extract F, and let I' be the resultant image after extracting F, the equation can be reduced to a form equivalent to the intrinsic image representation [1]

$$I'(x,y) = \beta(x,y)B(x,y) \tag{2}$$

which was solved using multiple images [14] and a single image [13]. By taking the advantage of the smooth β assumption, this paper takes an alternative approach to achieve better results in various problems in computational photography which requires a small amount of user interaction: Figure 1 shows an example of extracting a glass layer with substantial transparency. Note that the background is well separated from the glass layer. Figure 2 shows an extracted smooth shadow with a hard shadow boundary, indicating that both high and low frequency components coexist in the layer. Note that the textures of the image B are preserved after shadow removal.

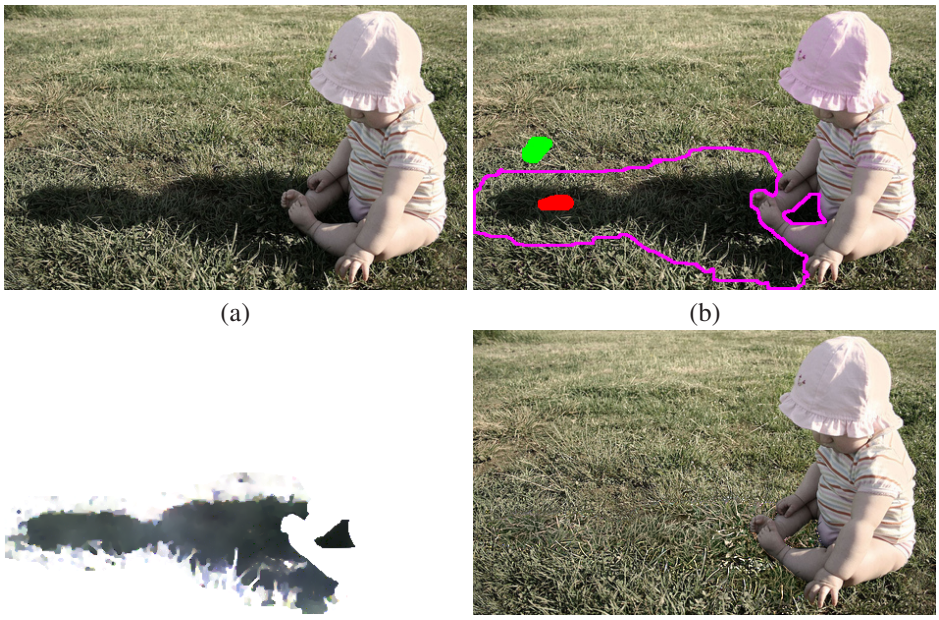
It turns out that the same algorithm we proposed, EM-HMM, can be used to extract both F and β. If β is not refractive, B can be simply obtained by I'/β. Related to our work is natural image matting, which has been used to extract, from a single image, overlapping layers with transparent boundaries. Figure 3 demonstrates that while current state-of-the-art matting techniques [3, 11, 5] can also be used to extract the martini glass, the F layers (shown here as the glass, highlight, and environment reflection layers combined) produced by our EM-HMM algorithm, which employs the hidden Markov model and considers first-order spatial neighborhood, is more homogeneous and less susceptible to the structure caused by the observed background. As we shall demonstrate, our method outputs a set of color labels per pixel which serves to reduce the inherent color ambiguities. When properly employed, we believe that such reduction should be very useful to image matting algorithms in general.

*This research was supported by the Research Grant Council, Hong Hong (CERG), under grant no. 620207.



(a) (b) (c) (d)

Figure 1. (a) Input image, (b) user scribbles for collecting relevant color cues, (c) transparent layer β, and (d) the background image B extracted by our method.



(a) (b) (c) (d)

Figure 2. (a) Input image. (b) Input strokes. (c) extracted shadow β, (d) the background B after removing the shadow. Note that the extracted transparent shadow is smooth, of spatially-varying intensity, and free of grass textures. The region delimited by pink loops are processing regions.

2. Overall Approach

Based on the above analysis, our algorithm steps are:

1. Extract F (section 3). We first extract overlapping layers which can be opaque or substantially transparent like the martini glass shown in Figure 3. Our EM-HMM algorithm produces a set of color labels at each pixel.

2. Extract β (section 4). Next, we solve Eqn. 2. We use the same EM-HMM algorithm to determine the amount a pixel is attenuated by the smooth β. Given that β is smooth and B contains most image gradients, we incorporate these considerations as optimization constraints in the Bayesian MAP estimation.

3. Extract F

We develop an EM-HMM algorithm to extract F. Our approach can be regarded as soft segmentation: for each pixel i, compute an optimal set of n soft labels, $\alpha_{ij} \in [0, 1]$, $\sum_j^n \alpha_{ij} = 1$ where n is the total number of color segments in the image. In fact, many natural image matting techniques can also be used if $n = 2$ and if F is largely opaque. In this paper, we take advantage of the smoothness assumption, so we can extract $n \geq 2$ overlapping color segments

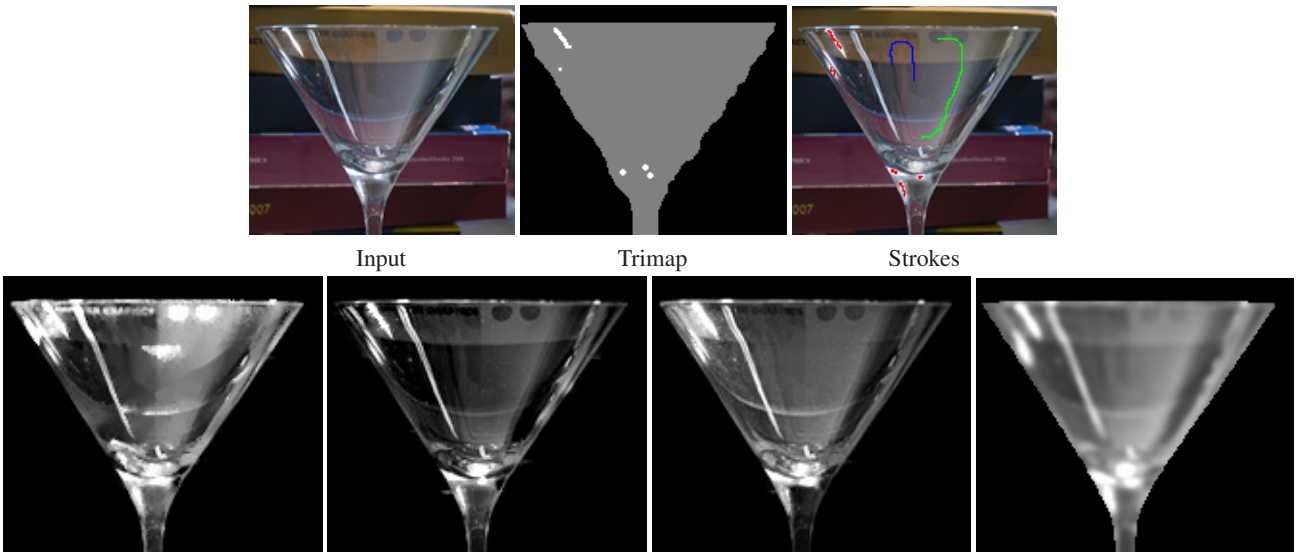


Figure 3. The input image, the trimap (used in Bayesian and Poisson matting), the input strokes (used in our method), and the F results of natural image matting using Bayesian matting, Poisson matting, the closed-form solution, and our method. For this example, our result is less susceptible to the background textures or structures.

which can be substantially transparent in front of a complex background.

The user-scribbled color samples (e.g. Figures 1–3) provide the necessary color constraints for our EM-HMM algorithm. Consider the two scenarios: 1) Suppose we know the expected color, then we can estimate the corresponding soft region label. 2) Suppose on the other hand the soft region label is known, it will help the color label estimation. In our case, both colors and soft labels are unknown. This becomes a "chicken-and-egg" problem. We propose to optimize the two variables by alternating optimization. EM algorithm, which is one form of alternating optimization guaranteed to converge [2], is a good choice of the algorithm.

The following EM derivations except the novel use of HMM are quite standard and should be familiar to readers with knowledge of EM (a gentle tutorial is available [2]).

By using a stroke-based interface to sample colors, we model the collected color statistics using Gaussians to allow uncertainty in subsequent estimation. In our formulation the three color channels can be processed individually or together. Both produce similar visual results, so for simplicity of notations, the following equations assume a single channel.

Using terminologies typical of EM formulations, let $\mathbf{O} = \{\{\mu_1, \sigma_1\}, \{\mu_2, \sigma_2\}, \cdots, \{\mu_n, \sigma_n\}\}$, $j = 1 \cdots n$, be the set of observations where μ_j and σ_j are respectively the mean and standard variation of the colors sampled inside region j. Let $\mathbf{R} = \{r_i\}$ be the set of hidden variables that describes the classes labels at all pixels. $r_i = j$ if pixel i belongs to region j. The cardinality of $\mathbf{R}$, $|\mathbf{R}|$, is equal to the total number of pixels N to be processed (e.g. the whole image or the pixels inside the object silhouette). The objective function is given by

$$\mathbf{\Theta}^* = \arg\max_{\mathbf{\Theta}} P(\mathbf{O}, \mathbf{R}|\mathbf{\Theta}) \tag{3}$$

where $P(\mathbf{O}, \mathbf{R}|\mathbf{\Theta})$ is the complete-data likelihood to be maximized and $\mathbf{\Theta} = \{c_i\}$ is a set of parameters to be estimated, where c_i is the expected color at pixel i. To estimate $\mathbf{\Theta}^*$, the EM algorithm computes the expected value of the complete-data log-likelihood $\log P(\mathbf{O}, \mathbf{R}|\mathbf{\Theta})$ with respect to $\mathbf{R}$ given the observation $\mathbf{O}$ and the current estimated parameter $\mathbf{\Theta}'$:

$$Q(\mathbf{\Theta}, \mathbf{\Theta}') = \sum_{\mathbf{R} \in \varphi} \log P(\mathbf{O}, \mathbf{R}|\mathbf{\Theta}) P(\mathbf{R}|\mathbf{O}, \mathbf{\Theta}') \tag{4}$$

where φ is the space containing all possible $\mathbf{R}$ with cardinality equal to N.

3.1. Expectation

We first define the marginal probability $p(\mathbf{O}|r_i, \mathbf{\Theta}')$ so that we can maximize the expectation Q defined by Eqn. 4 by proceeding to the next iteration given the current parameter estimation. If $r_i = j$, c_i should be similar to μ_j:

$$p(\mathbf{O}|r_i, \mathbf{\Theta}') \propto \begin{cases} \exp(-\frac{|c_i - \mu_1|^2}{2\sigma_1^2}) & \text{if } r_i = 1 \\ \vdots & \\ \exp(-\frac{|c_i - \mu_n|^2}{2\sigma_n^2}) & \text{if } r_i = n \end{cases} \tag{5}$$

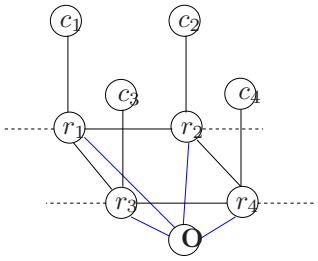


Figure 4. The HMM model for estimating the set of soft labels at each pixel.

Without any prior information, we let $p(r_i = j|\mathbf{\Theta}') = \frac{1}{n}$ be the mixture probability. Given $\mathbf{\Theta}'$ only, we have:

$$p(\mathbf{O}|\mathbf{\Theta}') \propto \frac{1}{n}\sum_j \exp(-\frac{|c_i - \mu_j|^2}{2\sigma_j^2}) \tag{6}$$

Let α_{ij} be the probability of pixel i belonging to region j, which is the output soft label we need. Then, $\alpha_{ij} = p(r_i = j|\mathbf{O}, \mathbf{\Theta}')$, or

$$\begin{aligned}\alpha_{ij} &= \frac{p(r_i = j, \mathbf{O}|\mathbf{\Theta}')}{p(\mathbf{O}|\mathbf{\Theta}')} = \frac{p(\mathbf{O}|r_i = j, \mathbf{\Theta}')p(r_i = j|\mathbf{\Theta}')}{p(\mathbf{O}|\mathbf{\Theta}')} \\ &= \exp(-\frac{|c_i - \mu_j|^2}{2\sigma_j^2}) / \sum_m \exp(-\frac{|c_i - \mu_m|^2}{2\sigma_m^2})\end{aligned} \tag{7}$$

3.2. Maximization

Given the marginal distribution α_{ij} estimated in the E-Step, we can maximize the likelihood in Eqn. 3 by optimizing the parameters in Eqn. 4 using the estimated α_{ij}. We make use of the assumption of smooth layers, and decompose $P(\mathbf{O}, \mathbf{R}|\mathbf{\Theta})$ into a combination of simple elements based on the Hidden Markov Model (HMM) assumptions: 1) The hidden variable r_i depends only on the hidden variables of its first-order-neighbors. 2) The observation at i depends only on the hidden variable at i. These assumptions incorporate the smoothness consideration. See Figure 4 for a graphical representation of the HMM assumptions. Therefore:

$$P(\mathbf{O}, \mathbf{R}|\mathbf{\Theta}) = \prod_i \prod_{k\in\mathcal{N}_i} p(r_i|r_k, \mathbf{\Theta})p(\mathbf{O}|r_i, \mathbf{\Theta}) \tag{8}$$

where $\mathcal{N}_i$ is a set of right and bottom neighbors of i and $p(r_i|r_k, \mathbf{\Theta}) = 1$ if k does not exist.

With Eqn. 8, $Q(\mathbf{\Theta}, \mathbf{\Theta}')$ in Eqn. 4 can be rewritten as:

$$\sum_{\mathbf{R}\in\varphi} \log P(\mathbf{O}, \mathbf{R}|\mathbf{\Theta})P(\mathbf{R}|\mathbf{O}, \mathbf{\Theta}') \tag{9}$$

$$\begin{aligned}&= \sum_{\mathbf{R}\in\varphi} \log\{\prod_i \prod_{k\in\mathcal{N}_i} p(r_i|r_k, \mathbf{\Theta})p(\mathbf{O}|r_i, \mathbf{\Theta})\}P(\mathbf{R}|\mathbf{O}, \mathbf{\Theta}') \\ &= \sum_{\mathbf{R}\in\varphi}\sum_i\sum_{k\in\mathcal{N}_i} \log\{p(r_i|r_k, \mathbf{\Theta})\}p(r_i, r_k|\mathbf{O}, \mathbf{\Theta}') \\ &\quad + \sum_{\mathbf{R}\in\varphi}\sum_i \log\{p(\mathbf{O}|r_i, \mathbf{\Theta})\}p(r_i|\mathbf{O}, \mathbf{\Theta}')\end{aligned} \tag{10}$$

Here, we have two terms to be defined: $p(r_i|r_k, \mathbf{\Theta})$ and $p(r_i, r_k|\mathbf{O}, \mathbf{\Theta}')$ (Note that $p(\mathbf{O}|r_i, \mathbf{\Theta})$ was defined similarly as Eqn. 5 and $p(r_i = j|\mathbf{O}, \mathbf{\Theta}') = \alpha_{ij}$).

For $p(r_i, r_k|\mathbf{O}, \mathbf{\Theta}')$ which models the first-order connection between two adjacent nodes with hidden variables r_i and r_k, regardless of the values of r_i and r_k, it equals to 1 because of the HMM assumptions (Figure 4).

For $p(r_i|r_k, \mathbf{\Theta})$, we assume that if the neighborhood pixel pair belongs to the same region, the expected color should be similar, i.e., if $r_i = r_k$, the color c_i should be similar to the color c_k and vice versa.

Suppose the noise model obeys the Gaussian distribution, $p(r_i|r_k, \mathbf{\Theta})$ can be modeled as:

$$p(r_i|r_k, \mathbf{\Theta}) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{|c_i - c_k|^2}{2\sigma^2}) \tag{11}$$

where σ^2 describes the variance of the region color. Since $r_i \in \{0, 1, \cdots, n\}$, we rewrite $Q(\mathbf{\Theta}, \mathbf{\Theta}')$ as

$$\begin{aligned}&Q(\mathbf{\Theta}, \mathbf{\Theta}') \\ &= \sum_{\mathbf{R}\in\varphi}\sum_i\sum_{k\in\mathcal{N}_i} \log\{p(r_i|r_k, \mathbf{\Theta})\}p(r_i, r_k|\mathbf{O}, \mathbf{\Theta}') \\ &\quad + \sum_i\sum_j \log\{p(\mathbf{O}|r_i = j, \mathbf{\Theta})\}\alpha_{ij}\end{aligned} \tag{12}$$

$$\begin{aligned}&= \sum_i\sum_{k\in\mathcal{N}_i} \log\left(\frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{|c_i - c_k|^2}{2\sigma^2})\right) \\ &+ \sum_i\sum_j \log\left(\frac{1}{\sigma_j\sqrt{2\pi}} \exp(-\frac{|c_i - \mu_j|^2}{2\sigma_j^2})\right)\alpha_{ij}\end{aligned} \tag{13}$$

where σ controls the measurement uncertainty (we simply set σ equals to the mean of all σ_j). To maximize Q, we differentiate Q w.r.t c_i and set the first derivative equal to zero to obtain the parameter updating rule:

$$c_i = \left(\sum_{k\in\mathcal{G}_i} c_k + \sigma^2\sum_j \frac{\alpha_{ij}}{\sigma_j^2}\mu_j\right) / \left(4 + \sigma^2\sum_j \frac{\alpha_{ij}}{\sigma_j^2}\right) \tag{14}$$

where $\mathcal{G}_i$ is first-order neighbors of pixel i. Hence, in the M-Step, the updating rule (compute c_i by Eqn. 14) is applied. The E-Step (compute α_{ij} by Eqn. 7) and M-Step are iterated alternately until convergence. The initial assignment of c_i is set as the pixel's color I_i.

4. Extract β

After extracting αs which represent the soft segmentation result for F, we extract β. The user marks up on the image to gather color samples in the background outside and inside of the transparent layer, where the two marked-up regions should have similar textures/structures.

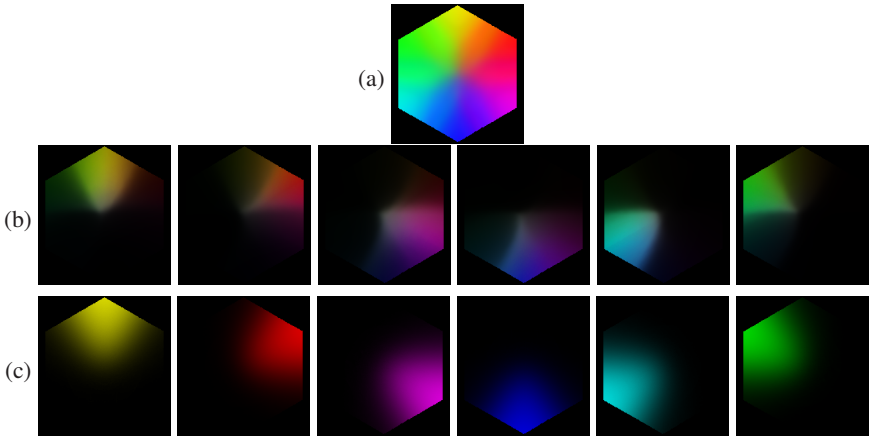


Figure 5. (a) Input synthetic image where the observed color of a pixel may be explained by a mixture of as many as six colors. The soft segments corresponding to (a) produced by using (b) our proposed method. (c) The ground truth soft segments. Note that the estimated soft segments are displayed by multiplying the estimated soft labels with the input image.

Let the two marked-up regions $\mathbf{O} = \{\{\mu_1, \sigma_1\}, \{\mu_2, \sigma_2\}\}$. By letting $n = 2$ the EM-HMM algorithm is used to compute the probability α_{ij}, i.e., $j = 0$ denotes unattenuated background, and $j = 1$ otherwise.

We modify the Bayesian MAP optimization in [15] to estimate β by incorporating αs to improve the results, where the optimal β is given by

$$\beta^* = \arg\max_{\beta} P(B^*|\beta)P(\beta) \tag{15}$$

where B^* is a rough estimation of the background without transparency attenuation. The estimation of B^* is the same as that of $\hat{\mathcal{F}}^*$ in section 3.3 of [15] after replacing the corresponding symbols, which solves a Poisson equation subject to a guidance field.

$P(B^*|\beta)$. To discern true image structures from image gradients caused by transparency attenuation, we use α to encode the probability of the observed image gradient at pixel x caused by attenuation. Let

$$m_{x,y} = \exp\left(-\frac{|\alpha_{x1} - \alpha_{y1}|^2}{2\sigma_m^2}\right) \tag{16}$$

where σ_m is the uncertainty in the smoothness measurement. By assuming the error distribution to be Gaussian, we define the likelihood $P(B^*|\beta)$ as

$$\exp\left(-\frac{\sum_{\{x,y\}\in\mathcal{R}} ||\nabla I'_{x,y} - m_{x,y}\nabla(\beta B^*)_{x,y}||^2}{2\sigma_1^2}\right) \tag{17}$$

which measures the fidelity between the image gradients of I' and the estimated βB^* weighted by m. $\{x, y\}$ are first-order neighbors in the valid processing region of I', denoted above by $\mathcal{R}$, obtained by masking out irrelevant regions by intelligent scissor and extracting F by EM-HMM. σ_1 is the standard deviation of the measurement error.

$P(\beta)$. By assuming the transparent object to be homogeneous, we use the following smoothness prior $P(\beta)$ weighted by $m_{x,y}$ as

$$\exp\left(-\frac{\sum_{\{x,y\}\in\mathcal{R}} m_{x,y}||\beta_x - \beta_y||^2}{2\sigma_2^2}\right) \tag{18}$$

where σ_2 is the uncertainty in the smoothness prior.

5. Results

For all examples in this paper, it takes less than a minute to process each case on a 2.8 GHz PC with 1G RAM.

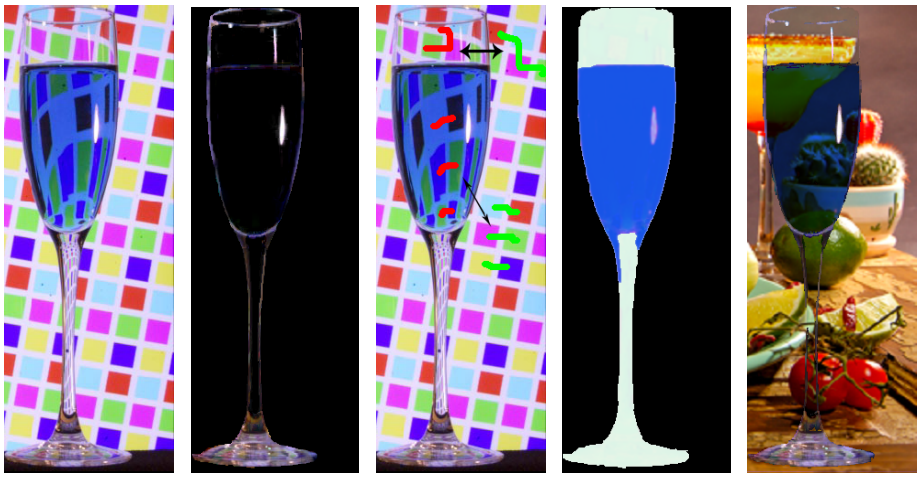
Synthetic case. Figure 5(a) shows a hexagon with spatially-varying colors, produced by compositing six soft color regions as shown. A single pixel's color can be explained by as many as six colors. Figure 5(b) shows the results produced by our method. (c) shows the ground truth soft segment of the synthetic image. To initialize the algorithms, we used the same set of mean colors estimated by k-means clustering ($k = 6$) for both of the methods. We calculated the difference of the estimated region label with the ground truth region label by $(\sum_i |\alpha_{ij} - \alpha_{ij}^G|)/N$, where α_{ij} is the estimated region label at i, α_{ij}^G is the corresponding ground truth. The mean difference from the six region for our method is 7.0 (scale 0-255).

Glass and Shadow removal. We demonstrated in Figure 1 glass removal, and in Figure 2 shadow removal. Shadowed (resp. glass) and unshadowed (resp. non-glass) color samples are collected and then input to our EM-HMM algorithm. The extracted smooth layer, which is of spatially-varying intensity and free of any textures, can be used in image and shadow matting, Figure 6.

Figure 7 shows a result of separating F and β, using example in [16]. For the results on F, we marked a stroke on

(a) (b)

Figure 6. The extracted transparent layers can be used in (a) image matting and (b) shadow matting.



Input Glass F Color markup smooth β Composite

Figure 7. Layer decomposition from a single image (input image from [16]). After extracting the highlight layer, the user marks up on the image attenuated and unattenuated background to gather color and texture statistics, in order for the system to optimize the smooth β free of any textures.

the highlight. The object boundary was obtained from the object silhouette, which is given by GrabCut [8] or intelligent scissors [7]. To extract the blue liquid, we first mask out the F. Then, in the "color markup", we indicate to the system the background colors before and after the attenuation, after which the Bayesian MAP estimation automatically produces the smooth β shown. The image composite is shown on the right. Note that we do not consider refraction.

Colorization. We perform soft color segmentation respectively in both the gray image and the example image, the latter of which provides the color statistics. Figure 8 shows our result. In this example, we show the effects when the color statistic σ_j in the observation $\mathbf{O}$ is adjusted by user to achieve different effects. In this example the $\mathbf{O}$'s in both the gray and example images are obtained automatically by K-mean clustering instead of user strokes. Here we input different σ_j as different initialization for the EM-HMM algorithm. In Figure 8(d) is the result with small σ_j, and Figure 8(g) shows the result with large σ_j. From the corresponding zoomed images shown in Figure 8(e), (f), (h) and (i), we can notice that the boundary of the result with small σ_j is quite unnatural and noticeable. This is because using small σ_j is analogous to a hard segmentation. The boundary of the result with large σ_j, on the other hand, is more natural. But a large σ_j makes the blue color in the sky overlap with the green color of the grass in the color space. So the blue color depicted in Figure 8(g) is not as pure as that in Figure 8(d). Figure 8(c) shows the result when global colorization is performed.

6. Conclusion

Layer separation from a single image is a massively ill-posed problem in its general form. This paper proposes to solve an easier but useful alternative, and presents an EM-HMM algorithm to separate smooth layers and the substantially-textured background from a single image. The EM alternately optimizes the soft label and the expected color at each pixel, where the HMM is used to maintain spatial coherency of the smooth layers. The image textures of the background layer are explicitly preserved by solving

Figure 8. Color transfer to a gray scale image. (a) The image to be colorized. (b) The example image, which provides the relevant color statistics for colorizing (a). (c) Result by global color transfer. (d) Result by our method, with small σ_j input for the sky and the grass regions. (e) and (f) Zoom in of respective layers in (d). (g) Result with large σ_j. (h) and (i) Zoom in of respective layers in (g). See the electronic version for color visualization.

the Bayesian MAP estimation problem. Our proposed algorithm is demonstrated to produce good results in various computational photography applications. The executable of our code will be available at the first author's homepage.

Acknowledgment

The authors would like to thank Ruonan Pu for her help in generating the results shown in Figure 3 and 7.

References

[1] H. Barrow and J. Tenenbaum. Recovering intrinsic scene characteristics from images. In *CVS78*, pages 3–26, 1978.

[2] J. Bilmes. A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical Report ICSI-TR-97-021, ICSI, 1997.

[3] Y.-Y. Chuang, B. Curless, D. H. Salesin, and R. Szeliski. A bayesian approach to digital matting. *CVPR'01*, pages 264–271, 2001.

[4] H. Farid and E. Adelson. Separating reflections from images by use of independent component analysis. 16(9):2136–2145, 1999.

[5] A. Levin, D. Lischinski, and Y. Weiss. A closed form solution to natural image matting. In *CVPR '06*, pages 61–68, 2006.

[6] A. Levin and Y. Weiss. User assisted separation of reflections from a single image using a sparsity prior. In *ECCV04*, volume I, pages 602–613, 2004.

[7] E. N. Mortensen and W. A. Barrett. Intelligent scissors for image composition. In *SIGGRAPH*, pages 191–198, 1995.

[8] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004.

[9] B. Sarel and M. Irani. Separating transparent layers of repetitive dynamic behaviors. In *ICCV05*, pages I: 26–32, 2005.

[10] Y. Schechner, N. Kiryati, and R. Basri. Separation of transparent layers using focus. *IJCV*, 39(1):25–39, August 2000.

[11] J. Sun, J. Jia, C.-K. Tang, and H.-Y. Shum. Poisson matting. *ACM Transactions on Graphics*, 23(3), 2004.

[12] R. Szeliksi, S. Avidan, and P. Anandan. Layer extraction from multiple images containing reflections and transparency. In *CVPR00*, volume 1, pages 246–253, 2000.

[13] M. Tappen, W. Freeman, and E. Adelson. Recovering intrinsic images from a single image. In *NIPS*, pages 1343–1350, 2002.

[14] Y. Weiss. Deriving intrinsic images from image sequences. In *ICCV01*, pages II: 68–75, 2001.

[15] T.-P. Wu and C.-K. Tang. A bayesian approach for shadow extraction from a single image. In *ICCV05*, pages I: 480–487, 2005.

[16] D. E. Zongker, D. M. Werner, B. Curless, and D. H. Salesin. Environment matting and compositing. In *SIGGRAPH '99*, pages 205–214, 1999.