A Matte-less, Variational Approach to Automatic Scene Compositing

Shanmuganathan Raman* and Subhasis Chaudhuri
Vision & Image Processing Lab, Department of Electrical Engineering, IIT Bombay, India
(shanmuga,sc)@ee.iitb.ac.in

Abstract

In this paper, we consider the problem of compositing a scene from multiple images. Multiple images, for example, can be obtained by varying the exposure of the camera, by changing the object at focus, or by simply sampling a video sequence at arbitrary time instants. We develop this problem in an optimization framework and then adopt a variational approach to derive a generalized algorithm which will be able to solve diverse applications depending on the nature of the input images. Our approach has distinct advantages over the existing digital compositing techniques, such as alpha matting and alpha blending, which require an explicit preparation of the matte while there is no such requirement in the proposed technique. We demonstrate the usefulness of our approach through results from diverse applications in computer vision.

1. Introduction

Digital compositing is a modern technique used extensively in applications where multiple input images contribute to generating the desired output image ([2],[19],[26]). Compositing is presently performed using mattes which fetch desired elements from each of the input images. Alpha matting and alpha blending are the commonly known approaches in which such mattes are employed. The matte is prepared separately, often manually, and is stored as a separate channel in each of the input images along with the intensity channel(s). The major overhead in these approaches is the preparation of the mattes using various tools by a skilled user.

Matte can be calculated from an image using a Bayesian approach [6], using iterative energy minimization with the help of Markov random field [11], by extending blue screen matting using some knowledge about boundary location [23], by solving Poisson equations with the help of user given boundary information [27], or by optimization approach based on belief propagation involving segmentation [29]. In video matting, matte is generated from a video sequence by learning the statistics of the fore(back)ground [1], by motion cue [5] or by separating the object layers through defocus [17].

In alpha blending, a parameter α which represents the matte information weighs each of the input images appropriately to produce the desired output image [25]. Let f be the desired output image and g_m, where $0 \leq m \leq K$ be the K input images. The prepared matte is given by $\alpha_m(x, y)$. The alpha blending, which requires explicit computation of α_m, performs the following operation:

$$f(x,y) = \sum_{m=1}^{K} \alpha_m(x,y) g_m(x,y), \sum_{m=1}^{K} \alpha_m(x,y) = 1. \quad (1)$$

In this paper, we adopt a matte-less approach for scene compositing which does not require explicit computation of the mattes. We pose digital compositing as an unconstrained optimization problem involving selection of locally high contrast pixels yet maintaining a certain smoothness over the neighborhood. We adopt a variational approach to solve the optimization problem and derive a generalized algorithm for solving the digital compositing problem.

It is interesting to note that the problem of compositing has always been considered as an art, rather than engineering. Although the industry has made use of various technologies such as blue screen photography, rotoscoping, polyfill (polygon filling) method, etc., very little effort has gone into automating the entire process of compositing except those cited earlier. Some recent trends include separation of object layers [28] or video object planes from different input videos for matte preparation [18]. However such methods would require explicit computation of structure from motion and the accuracy of the matte depends on the accuracy with which the layers can be separated. Our method does not suffer from such a problem. We do not require computation of image statistics like those in ([1], [6]) or scene modeling as in ([11],[27]) notwithstanding the fact that the proposed method is a data dependent process.

We apply our algorithm to solve diverse applications related to digital compositing in computer vision such as high dynamic range (HDR) imaging problem, recovering pinhole

*This work was partly supported by Microsoft Corporation and Microsoft Research India under the Microsoft Research India PhD Fellowship Award.

equivalent image from multiple defocused images, generation of random texture from multiple different texture images and motion trail generation from a video sequence. We show that excellent compositing can be done in all cases using the same algorithm.

2. Proposed Algorithm

In this section, we develop an algorithm which performs compositing on multiple input images of the scene. Let the algorithm take K images as input. Let $g_m(x, y)$ be the intensity value of the pixel at location (x, y) of the mth image, where $1 \leq m \leq K$. Let $f(x, y)$ be the unknown image to be constructed from $g_m(x, y)$.

2.1. Possible Formulations

General compositing problem can be formulated as

$$f(x,y) = \sum_{m=1}^{K} w_m(x,y) g_m(x,y) \tag{2}$$

where $w_m(x, y)$ is our representation for weighting function which replaces $\alpha_m(x, y)$ in alpha blending. As a simple case, we can model $w_m(x, y)$ as

$$w_m(x,y) = \begin{cases} 1 & for\ m = k\ (some\ index) \\ 0 & otherwise \end{cases} \tag{3}$$

Here our objective is to select a value for a pixel from a specific observation and it is a pixel-based approach. This approach is carried out by many artists in computer graphics. In some cases, this approach is carried out automatically, when this becomes a problem of combinatorial optimization. Some specific tasks in this paper require pixel-wise compositing as will be shown later in sub-sections 3.1-3.3.

Alternatively, we can select $w_m(x, y)$ using a region based approach as

$$w_m(x,y) = \begin{cases} 1 & for\ (x,y)\ \epsilon\ R_m \\ 0 & otherwise \end{cases} \tag{4}$$

where, R_m is the distinct region to be selected from the mth input image. The layer based approach explained in section 1 uses such a matte. Application defined in subsection 3.4 requires such a matte.

2.2. Matte-less, Variational Approach

To illustrate and to motivate, we consider the HDR problem. We require multiple, differently exposed images $g_m(x, y)$ as the input images. We want to design an algorithm to produce the desired HDR-like image $f(x, y)$ which is similar to the corresponding HDR image generated using standard HDR techniques ([16],[21],[22]). We shall now design our weighting function $w_m(x, y)$ with regard to this application. We want the selected pixel to have a high local contrast, yet blending smoothly across different regions while compositing from different images. Therefore we require a variational approach to solve this. The basic formulation of this problem can be designed as

$$f(x,y) = arg\min_f \Bigg[\int_x \int_y \Bigg\{ \Bigg(f(x,y) - \sum_{m=1}^{K} w_m(x,y) W_m\big(g_m(x,y)\big) \Bigg)^2 + \lambda \Big(f_x^2 + f_y^2 \Big) \Bigg\} dx\, dy \Bigg] \tag{5}$$

where λ is a regularization parameter which appropriately weighs the smoothness term $(f_x^2 + f_y^2)$, applied everywhere and W_m is the warping function which compensates for the motion of camera corresponding to the m th input image [32]. If we assume that the warping has already been carried out (note that this may require camera calibration) and incorporated into the input images $g_m(x, y)$, the basic formulation of the problem can be given as

$$f(x,y) = arg\min_f \Bigg[\int_x \int_y \Bigg\{ \Bigg(f(x,y) - \sum_{m=1}^{K} w_m(x,y) g_m(x,y) \Bigg)^2 + \lambda \Big(f_x^2 + f_y^2 \Big) \Bigg\} dx\, dy \Bigg] \tag{6}$$

Our main objective in the development of this algorithm is to model the weighting function $w_m(x, y)$ which weighs each of the pixels of the input images $g_m(x, y)$ in such a way so that it is amenable to an iterative solution, avoiding a combinatorial search, while w_m is still a data dependent term.

One easy way to model $w_m(x, y)$ is to pick one pixel from K images corresponding to the pixel location (x, y) using Equation (3) or (4). So, $w_m(x, y)$ takes one of the two values in $\{0, 1\}$ at any location (x, y). This leads us to a combinatorial problem and the variational framework has no proper solution. This cannot be solved using the Euler-Lagrange formulation which offers an iterative solution. Hence, we relax the condition on the weighting function $w_m(x, y)$ allowing it to take real values in the range $[0, 1]$ subject to the constraint

$$\sum_{m=1}^{K} w_m(x,y) \;=\; 1. \tag{7}$$

We propose a weighting function which would optimally weigh pixels of each of the K images so that the unknown image $f(x, y)$ is very close to the desired image with uniform illumination or contrast. The new weighting function is as shown below.

$$w_m(x,y) = \frac{B_m(x,y)}{A(x,y)} \tag{8}$$

where

$$B_m(x,y) = \big(C + \sigma_m^2(x,y)\big)\big(f(x,y) - g_m(x,y)\big)^2 \quad (9)$$

$$A(x,y) = \sum_{n=1}^{K} B_n(x,y) \quad (10)$$

where $\sigma_m^2(x,y)$ is the local variance around the pixel location (x,y) in the mth image, C is a real number meant to vary the influence of $\sigma_m^2(x,y)$ on the weighting function as well as to prevent the condition of w_m being zero in all images at homogeneous regions, and $(f(x,y) - g_m(x,y))^2$ is a measure of how the composited image is different from the mth observation at a given location. The division by the sum over all the input images is done to ensure that the weights assigned to K images sum to unity for a given pixel (x,y). Similar weighting functions have been employed in colorization [15] and image segmentation ([24],[31]).

The choice of the above weighting function w_m stems from the fact that if the pixel in a particular observation has a high local variance σ_m^2, then it should have a higher weight while compositing. Similarly, we want to emphasize selection of the over-exposed view at an under-illuminated pixel or the under-exposed view at an over-illuminated region. This is achieved through the choice of the second term $(f - g_m)^2$. The composited image would be away from being over- or under-exposed. This particular choice of the weighting function has an additional mathematical justification. If the K input images correspond to the same scene with different noisy (assumed Gaussian) observations, then the solution would be an optimal noise smoothing filter.

This optimization problem is solved using calculus of variations using the corresponding Euler-Lagrange equation [12]. The iterative discretized version of the solution is as shown below.

$$f^{p+1}(i,j) = \overline{f}^{p}(i,j) - \frac{1}{\lambda}\Bigg[\Bigg\{f^p(i,j) - \frac{\sum_{m=1}^{K} B_m^p(i,j) g_m(i,j)}{A^p(i,j)}\Bigg\}\Bigg\{1 - \sum_{m=1}^{K} g_m(i,j)\left(\frac{A^p(i,j)D_m^p(i,j) - B_m^p(i,j)E^p(i,j)}{(A^p(i,j))^2}\right)\Bigg\}\Bigg] \quad (11)$$

where

$$D_m(i,j) = 2\big[(C + \sigma_m^2(i,j))(f(i,j) - g_m(i,j))\big] \quad (12)$$

$$E(i,j) = \sum_{n=1}^{K} D_n(i,j). \quad (13)$$

The suffix 'p' denotes the value of the variable at the pth iteration and $\overline{f}$ denotes the average value of f over its nearest 4-connected neighborhood.

3. Applications

In this section, we shall see how the algorithm we just developed can be applied to various applications. No modification of the proposed algorithm is required for any of these suggested applications. We start with the HDR application considering which the algorithm was developed.

3.1. HDR Imaging Problem

HDR imaging technique produces an image which has a higher dynamic range from multiple differently exposed low dynamic range (LDR) images ([16],[21],[22]). The HDR image, generated using traditional approaches, will be able to represent objects in both the brightly and poorly illuminated regions in the scene equally well. The problems with these approaches are that the generated HDR image can be displayed using specialized displays only and it usually occupies a larger memory. Also, it requires the camera response function (CRF) to be computed from the input images [9] and, therefore, it is computationally expensive. HDR image can be displayed in conventional displays such as monitors after performing a process called tone mapping [14].

Our algorithm generates the output image which has the same dynamic range as that of the input images. This image will still be able to represent objects in both the brightly and poorly illuminated regions equally well in spite of its limited dynamic range. The proposed method implicitly replaces the over- or under-exposed regions by their appropriately exposed views. Further, the image generated using this approach neither requires the computation of CRF nor needs a specialized hardware to display it. This image requires less storage space compared to HDR image due to its limited dynamic range. Further, our method does not require any knowledge about the camera exposure settings.

3.2. Generation of Pin-hole Image

Depth from defocus ([3],[7]) is an active area of research where two or more different observations are used to recover the dense depth map. One can use a suitable restoration technique (by measuring the relative defocus) [20] to recover the corresponding pin-hole image. On the other hand, if one is given a large number of such defocused observations along with individual camera parameter settings, the corresponding depth recovery technique is called depth from focus [13]. We show that the proposed technique can be used to combine these observations to obtain the pin-hole equivalent image very efficiently.

Consider multiple defocused images of a scene with each image having different parts of the scene in focus. Our algorithm can combine these images to produce an image which has all parts of the scene in focus. The choice of w_m in Equation (8) will enable us to pick up the pixel intensity at a point from that particular observation where the point is

in focus so that it has the maximum local contrast.

3.3. Random Texture Generation

In certain applications, such as template preparation in textile industry, it is often desired to combine different existing textures to generate a new random texture. Our algorithm can also be employed to generate a randomly textured image from multiple different textures. The nature of the output image texture is controlled by varying the parameter C in Equation (9). This varies the influence of the local variance $\sigma_m^2(x, y)$ on the output image texture. The composited image will have the appearance of having picked up small regions from different constituent textures in a repetitive manner while picking locally high contrast pixels from the constituent images.

3.4. Motion Trail Generation

Motion trail generation from a video sequence is widely used in motion based video indexing and retrieval [4], in video authentication [30] and in static representation of moving objects in standard comics.

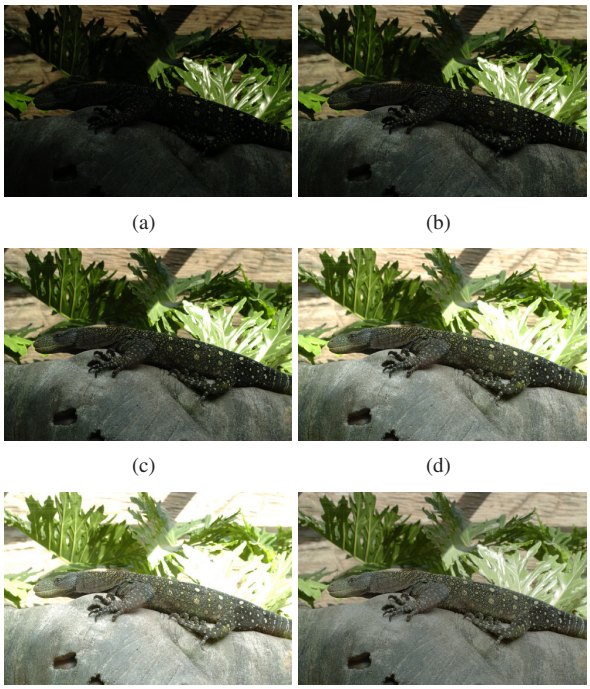
Consider the motion of an object in a video sequence. We want to represent the motion trail of the object in a single image. Our algorithm does this task very well by compositing several frames of the video sequence. We assume that the motion warping has already been carried out for the input frames if the camera is also in motion. This was already shown in Equations (5),(6). For a static region, the method will perform simple noise filtering. Whenever an object of interest is in motion, the corresponding edges (pixels with high local contrast) move along a motion and they get copied at a regular interval in the composited image.

4. Experimental Results

We have experimented on the suitability of the proposed algorithm on a number of practical applications. For reasons of brevity we show experimental results for only four specific applications that are highly relevant currently. We have used exactly the same computer code in all applications while maintaining the same value of the regularization parameter λ (used 10 in this study). Typical value of the parameter C used in our study is 100.

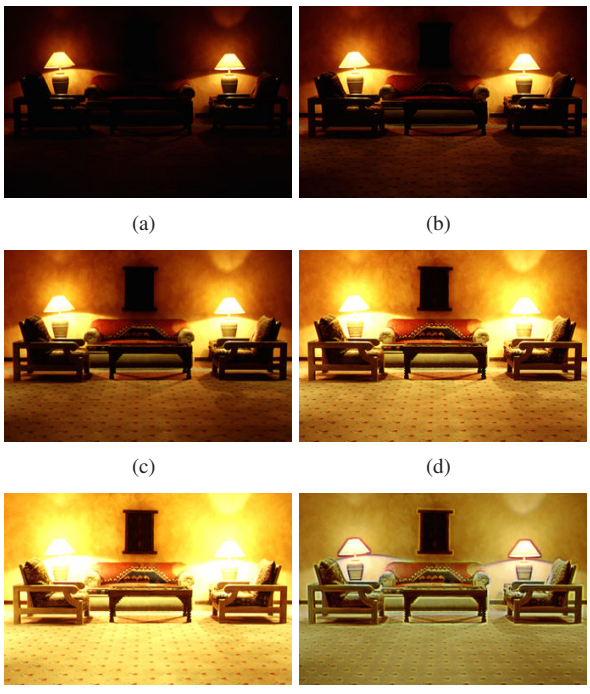
4.1. HDR Problem

In Figure 1(a-e), we show five different observations of a scene having a large variation in illumination. Each observation corresponds to different exposure settings (progressively more from (a) to (e)). Figure 1(f) shows the result of compositing after applying Equation (11). One can observe that the final output LDR image is uniformly illuminated and has similar characteristics of an HDR image. We call this output image as HDR-like image. Another example is shown in Figure 2 and the inference is quite similar.



(a) (b) (c) (d) (e) (f)

Figure 1. (a-e) Differently exposed images given as input, and (f) composited HDR-like image. *Data Courtesy:* Erik Reinhard, University of Bristol.

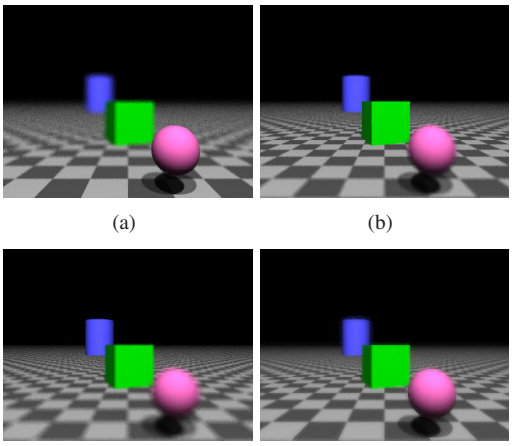


(a) (b) (c) (d) (e) (f)

Figure 2. (a-e) Differently exposed input images of another scene, and (f) composited HDR-like image. *Data Courtesy:* CAVE Lab, Columbia University.

4.2. Generation of Pin-hole Image

The images in Figures 3(a-c) (synthesized using Povray) of the scene have three objects (a ball, a cuboid, and a cylinder). Each of these input images have one of the objects in focus. These images are composited by our algorithm in Equation (11) and finally we get an image in Figure 3(d) in which all the three objects are in focus (pin-hole equivalent). In Figures 4(a-c), we have three defocused images of

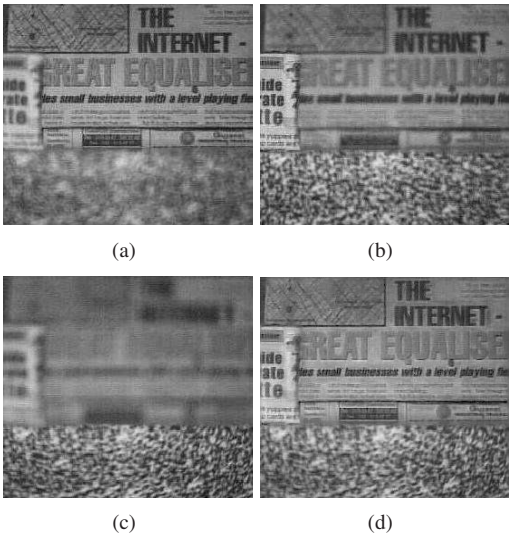


(a) (b)

(c) (d)

Figure 3. Input images with (a) front object (ball) in focus, (b) middle object (cuboid) in focus, (c) back object (cylinder) in focus, and (d) composited output image with all objects in focus.

a real world scene as the input images. The scene consists of planer surfaces at three distinct distances. Different layer in the scene is focused in each of the input images in 4(a-c). Our algorithm in Equation (11) works on these images and generates an output image in which all parts of the scene are properly in focus as shown in Figure 4(d).



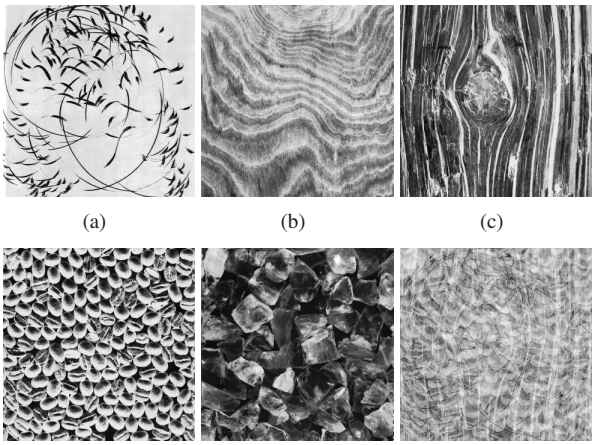
(a) (b)

(c) (d)

Figure 4. (a-c) Defocused input images with different depth layer in focus, and (d) composited pin-hole equivalent image.

4.3. Random Texture Generation

Figures 5(a-e) show a set of texture images used as input. Figure 5(f) shows the composited output from five input images using Equation (11). One does see a certain correlation in the composited image with each of the five input images. In some sense, one can relate this application to the problem of cross-dissolve [10] in graphics.



(a) (b) (c)

(d) (e) (f)

Figure 5. (a-e) Different texture images as input, and (f) composited new texture. *Data Courtesy:* Brodatz Texture Database.

4.4. Motion Trail Generation from a Video

(a) (b)

(c) (d)

Figure 6. (a-c) Selected frames (1,5 and 8) of a surveillance camera, and (d) composited image showing motion trail.

Consider frames of a video sequence captured using a surveillance camera. Figures 6(a-c) show selected frames of a video sequence in which two objects (a cycle and a car) are in motion. Eight such successive frames are used as input images. We apply our algorithm in Equation (11) on these eight frames, last frame being given more weight. Figure 6(d) shows the composited output image which depicts the motion trail of both the moving objects while other parts

of the scene are left unaltered. Since the cycle is moving slower than the car, the corresponding trail is smaller.

5. Conclusions

We have considered the problem of automatic compositing of images and proposed a matte-less, variational approach by selecting an appropriate weighting function. The novelty of the weighting function lies in the fact that it is completely data dependent. Hence our approach does not require any prior matte information. We have solved the problem adopting an optimization framework and arrived at an iterative solution. The iterative algorithm has been applied to diverse applications in computer vision and has been shown to perform very well in all cases. Our approach requires neither any user intervention nor pre-processing at any stage and is found to converge to optimal solution within 4-5 iterations in all cases. Further, if one requires to use a discontinuity preserving smoothness term [8], one can easily modify Equation 5 accordingly.

References

[1] N. Apostoloff and A. Fitzgibbon. Bayesian video matting using learnt image priors. In *CVPR*, volume 1, pages 407–414, Washington, DC, USA, 2004.

[2] J. F. Blinn. Compositing, part 1: Theory. *IEEE Computer Graphics & Applications*, 14(5):83–87, 1994.

[3] S. Chaudhuri and A. N. Rajagopalan. *Depth from Defocus: A Real Aperture Imaging Approach*. Springer-Verlag, New York, 1999.

[4] H. Chiou-Ting and T. Shang-Ju. Motion trajectory based video indexing and retrieval. In *ICIP*, volume 1, pages 605–608, Rochester, New York, USA, 2002.

[5] Y. Chuang, A. Agarwala, B. Curless, D. H. Salesin, and R. Szeliski. Video matting of complex scenes. In *SIGGRAPH*, pages 243–248, San Antonio, USA, 2002.

[6] Y. Chuang, B. Curless, D. H. Salesin, and R. Szeliski. A bayesian approach to digital matting. In *CVPR*, volume 2, pages 264–271, Kauai Marriott, Hawaii, 2001.

[7] P. Favaro and S. Soatto. *3-D Shape Estimation and Image Restoration*. Springer, 2006.

[8] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE PAMI*, 6:721–741, 1984.

[9] M. D. Grossberg and S. K. Nayar. Determining the camera response from images: what is knowable? *IEEE PAMI*, 25(11):1455–1467, 2003.

[10] M. Grundland, R. Vohra, G. Williams, and N. A. Dodgson. Cross dissolve without cross fade: Preserving contrast, color and salience in image compositing. *Eurographics*, 25(3):577–586, 2006.

[11] Y. Guan, W. Chen, X. Liang, Z. Ding, and Q. Peng. Easy matting - a stroke based approach for continuous image matting. *Eurographics*, 25(3):567–576, 2006.

[12] B. K. P. Horn. *Robot Vision*. The MIT Press, 1986.

[13] E. Krotkov. *Active Computer Vision by Cooperative Focus and Stereo*. Springer-Verlag, New York, 1989.

[14] G. W. Larson, H. Rushmeier, and C. Piatko. A visibility matching tone reproduction operator for high dynamic range scenes. *IEEE Trans. on Visualization and Computer Graphics*, 3(4):291–306, 1997.

[15] A. Levin, D. Lischinski, and Y. Weiss. Colorization using optimization. In *SIGGRAPH*, pages 689–694, Los Angeles, USA, 2004.

[16] S. Mann and R. W. Picard. On being undigital with digital cameras: extending dynamic range by combining exposed pictures. In *In Proc. of IS & T 48th annual conference*, pages 422–428, Cambridge, Massachusetts, USA, 1995.

[17] M. McGuire, W. Matusik, H. Pfister, J. F. Hughes, and F. Durand. Defocus video matting. In *SIGGRAPH*, pages 567–576, Los Angeles, USA, 2005.

[18] H. Nicolas and F. Denoual. Semi-automatic modifications of video object trajectories for video compositing applications. *Signal Processing*, 85(10):1970–1983, 2005.

[19] T. Porter and T. Duff. Compositing digital images. In *SIGGRAPH*, pages 253–259, Minneapolis, USA, 1984.

[20] A. N. Rajagopalan and S. Chaudhuri. An MRF Model-Based Approach to Simultaneous Recovery of Depth and Restoration from Defocused Images. *IEEE PAMI*, 21(7):577–589, 1999.

[21] E. Reinhard, G. Ward, S. Pattanaik, and P. Debevec. *High Dynamic Range Imaging: Acquisition, Display and Image-Based Lighting*. Morgan Kaufmann Publishers, 2005.

[22] M. A. Robertson, S. Borman, and R. L. Stevenson. Estimation-theoretic approach to dynamic range improvement using multiple exposures. *Journal of Electronic Imaging*, 12(2):219–228, 2003.

[23] M. Ruzon and C. Tomasi. Alpha estimation in natural images. In *CVPR*, volume 1, pages 18–25, Hilton Head Island, South Carolina, USA, 2000.

[24] J. Shi and J. Malik. Normalized cuts and image segmentation. In *CVPR*, pages 731–737, San Juan, Puerto Rico, 1997.

[25] A. Smith. Alpha and the history of digital compositing. *Microsoft Tech Memo 7*, 1995.

[26] P. Soille. Morphological image compositing. *IEEE PAMI*, 28(5):673–683, 2006.

[27] J. Sun, J. Jia, C. Tang, and H. Shum. Poisson matting. In *SIGGRAPH*, pages 315–321, Los Angeles, USA, 2004.

[28] H. Tao, H. S. Sawhney, and R. Kumar. Object tracking with bayesian estimation of dynamic layer representations. *IEEE PAMI*, 24(1):75–89, 2002.

[29] J. Wang and M. F. Cohen. An iterative optimization approach for unified image segmentation and matting. In *ICCV*, pages 936–943, Beijing, China, 2005.

[30] Y. Wei-Qi and M. S. Kankanhalli. Motion trajectory based video authentication. In *Proc. International Symposium on Circuits and Systems*, volume 3, pages 810–813, Bangkok, Thailand, 2003.

[31] Y. Weiss. Segmentation using eigenvectors: A unifying view. In *ICCV*, pages 975–982, Kerkyra, Corfu, Greece, 1999.

[32] G. Wolberg. *Digital Image Warping*. Wiley-IEEE Computer Society Press, 1990.