PHOTOMONTAGE FOR ROBUST HDR IMAGING WITH HAND-HELD CAMERAS

Ru Li, Xiaowu He, Shuaicheng Liu, Guanghui Liu, Bing Zeng

Institute of Image Processing
University of Electronic Science and Technology of China

ABSTRACT

This paper studies the image fusion from multiple images taken by hand-held cameras with different exposures. The existing methods often generate unsatisfactory results, such as the blurring/ghosting artifacts due to the problematic handling of camera motions, dynamic contents, and inappropriate fusion of local regions (e.g., over or under exposed). They often require high quality image registration before fusion. However, the accurate alignment is hard to obtain in many scenarios, such as scenes with large depth variations and dynamic textures. Besides, high quality alignment is also time consuming. In this paper, we only enable a rough registration by a single homography and combine the inputs seamlessly to hide any possible misalignment. Specifically, we propose to use a Markov Random Filed (MRF) function for the labelling of all pixels, which assigns different labels to different aligned input images. During the labelling, we choose well-exposured regions and skip moving objects simultaneously. Then, we combine a Laplace image according to the labels and construct the fusion result by solving the Poisson equation. We present various challenging examples to demonstrate the effectiveness and practicability of our approach.

Index Terms— Multi-exposure fusion, MRF, rough registration

1. INTRODUCTION

High-dynamic-range (HDR) imaging techniques have received lots of attentions from both research and industry communities. There are two main categories to do the synthesis. The first one directly synthesizes the result from multi-exposure images [1, 2], the other reconstructs an HDR image firstly and then applies the tone mapping for the display [3, 4]. Our method belongs to the first category.

Although the multi-exposure fusion (MEF) approaches have been studied extensively, there are still some drawbacks. For instance, ghosting/blurring artifacts are unavoidable. The merging techniques have been employed in many existing methods assumes that multiple exposure images are *accurately* aligned [5, 6, 7]. Thus, any misalignments due to either camera motions or dynamic contents will lead to ghosting/blurring artifacts. A Laplacian pyramid reconstruction scheme for image fusion was proposed in [8], which has been widely adopted in many follow-up works, e.g., [1, 2]. Specifically, in [2], for each pixel location, every aligned candidate pixel in the stack contributes to the final pixel value. If there are any misaligned regions, the fused results would suffer from the ghosting or blurring artifacts. Figure 1 shows some examples, where the input images are captured by a static camera, but the scenes contain dynamic objects (tree leaves in the left example and moving persons in the right example). The fused results by [2] suffer from the blurry (Fig. 1 left) and the ghosting (Fig. 1 right).

Later, some deghosting methods are proposed to handle the problems [9]. The methods based on energy optimization [10, 11]are introduced to maintain the image consistency or distinguish different parameters. Some patch-based methods [12, 13, 14] are proposed to handle inputs by patch level. However, the patch-based reconstruction is not always robust in some complicated situations, especially when encountered with dynamic textures (e.g., fountains, tree leaves in the wind), or structured regions. Fig. 2 shows such an example with much detailed information in the tree crown regions, where the methods of [12], [13] and [14] generate unsatisfactory results which lose many details.

In this work, instead of pursuing a fully registration, we propose to align the inputs roughly. The high quality fully registration is challenging. It is difficult to align images under different appearances [15, 16]. The large foreground [17] and near-range objects [18] further complicate the alignments. The non-parametric approaches such as optical flow tend to generate errors at regions with discontinuous depth [19, 20]. The patch-based reconstruction is also prone to produce errors as shown in Fig. 2. To pursue a robust solution, we abandon the requirement of fully registration and replace it by a rough registration with a single homography. We borrow the similar idea from [21] to composite the roughly aligned multi-exposure images. We select good exposed regions from the roughly aligned images and stitch them seamlessly. Our system can tolerate these alignment errors.

In summary, we focus on fusing images captured by hand-held cameras and do not require the high quality registration before fusion. To achieve this, we propose a solution that selects sub-image regions from different roughly aligned exposures by a MRF labelling and combine them seamlessly in

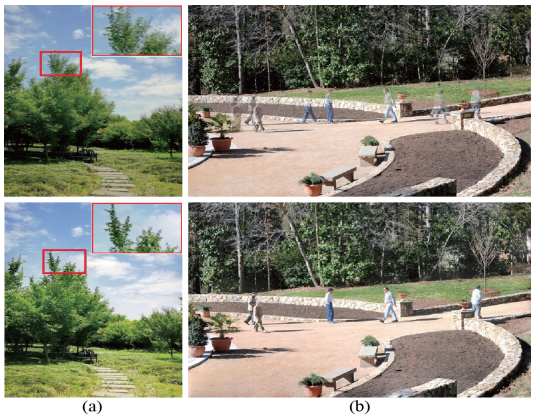


Fig. 1. Top images in (a), (b) are Mertens's results [2]. Our results are shown in the bottom. The comparison indicates that our method can effectively solve blur and dynamic objects

the gradient domain. In this way, each pixel value belongs to a single image such that maintaining details well and handling blurring effectively. Moreover, we consider the dynamic identification and exposure selection in the MRF optimization simultaneously. The selected regions are not only well-exposed but are also free from the interferences of dynamic objects/textures. Furthermore, we gather a dataset that consists of 120 group of images, ranging from daytime-nighttime, static-dynamic, and outdoor-indoor. We evaluate our method both qualitatively and quantitatively. The experiments demonstrate the effectiveness and robustness of our approach.

2. OUR METHOD

The input images are captured by hand-held cameras. The first step is to align them. We pick the image with medium exposure as the target. Specifically, we choose the Features from Accelerated Segment Test (FAST) [22] for the feature detection and track them by the Kanade-Lucas-Tomasi (KLT) [23]. We use the similar strategy as described in [24] for the feature prunning and boosting, as such features are more robust against the luminance differences. Figure 3 shows our system pipeline after the alignment. Without loss of generality, we take four input images as an example. Fig. 3(a) shows the aligned input images. Fig. 3(b) displays corresponding weight maps calculated by method [2]. Then, we use these weights to produce a label map (Fig. 3(c)) through MRF energy minimization. We collect the Laplace values at each pixel from different images according to the label map to yield a Laplace image. By solving the Poisson equation properly, we obtain the final result as shown in Fig. 3(d).

We optimize the following energy for the labelling:

$$E(X) = \sum_{i \in \upsilon} E_1(x_i) + \lambda' \sum_{i \in \upsilon} E_2(x_i) + \lambda'' \sum_{(i,j) \in \varepsilon} E_3(x_i, x_j) \quad (1)$$

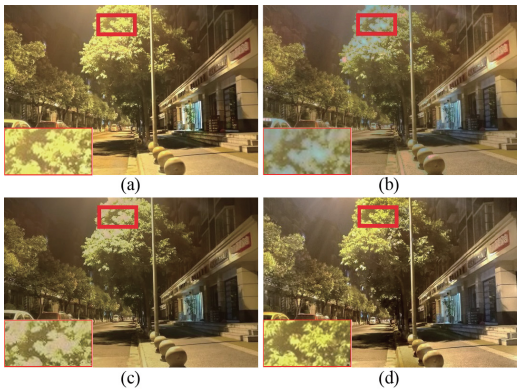


Fig. 2. (a) Result by [12]. (b) Result by Hu [13]. (c) Result by Sen [14]. (d) Our result.

where each candidate image corresponds to a label and x_i is the label of the pixel i. $E_1(x_i)$ and $E_2(x_i)$ are data terms, in which $E_1(x_i)$ is the likelihood energy representing exposure qualities. It encodes the color similarity of a pixel, indicating that which image it belongs to. $E_2(x_i)$ encodes the dynamic information. $E_3(x_i, x_j)$ is the smoothness term. υ is the set of all pixels and ε is the set of adjacent pixels. λ' and λ'' balance the terms. We set $\lambda' = 3$ and $\lambda'' = 5$ in our implementation. The energy can be minimized by Graph-cut [25].

2.1. Exposure Weights

$E_1(x_i)$ represents the pixel quality. It consists of three parts: *contrast*, *saturation* and *exposedness*, which are combined to form a weight map W for the fusion [2]. Here, we use the weight map as the probability for selecting image regions:

$$E_1(x_i = label) = \frac{1}{W_{label}(i) + eps} \quad (2)$$

where $label$ corresponds to the image labels; eps is set as 0.001 in our method to avoid $W_{label}(i) = 0$; W is the combined weight map of input image:

$$W_i = C_i \cdot S_i \cdot E_i, \quad (3)$$

where C_i, S_i, and E_i refer to the weight of *contrast*, *saturation*, and *exposedness*, respectively; "$\cdot$" is the Hadamard product; W is normalized between $(0, 1)$.

2.2. Dynamic Exclusion

We want to exclude the dynamic areas which need to first locate these regions. To achieve this, we adopt the approach [26], which applies an energy optimization to detect dynamic objects. The pixels of each input image are identified by a mask M:

$$M(i) = \begin{cases} 0 & i \in static\ areas \\ 1 & i \in dynamic\ areas \end{cases} \quad (4)$$

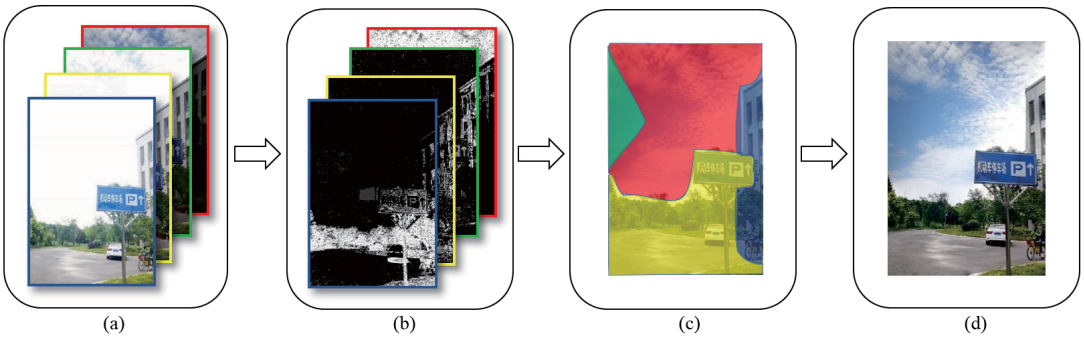


Fig. 3. The pipeline of our method. (a) Aligned input images where the third image is the reference. (b) Weights maps. (c) Final labels obtained by weight maps, with different color representing different input images. (d) The fusion result.

Then, M is feeded into $E_2(x_i)$:

$$E_2(x_i = label) = \begin{cases} \infty, & M(i) = 1 \\ 0, & M(i) = 0 \end{cases} \tag{5}$$

We increase the energy to avoid dynamic pixels. When a pixel is static, $E_2(x_i)$ does not introduce any penalties. Otherwise, if the pixel i of one input is detected as dynamic pixel, we set $E_2 = \infty$ to choose the value of pixel i from other images.

2.3. Spatial Smoothness

$E_3(x_i, x_j)$ is the smoothness term. It is defined as [27],

$$E_3(x_i, x_j) = |x_i - x_j| \cdot g(C_{ij}), \tag{6}$$

where $g(C_{ij}) = \frac{1}{1+C_{ij}}$ and $C_{ij} = ||C(i) - C(j)||^2$. $C(i)$ represents color information:

$$C(i) = [R(i)]^2 + [B(i)]^2 + [G(i)]^2 \tag{7}$$

where R, G and B are three channels of input image. C_{ij} is the $L2$-norm of the RGB color difference of two pixel i and j. Therefore, if two pixels have large difference, $g(C_{ij})$ is near to 0. $E_3(x_i, x_j)$ is a penalty term when adjacent terms are assigned with different labels.

Figure 4 shows the results of our labeling. The result of Fig. 4(b) is obtained by removing E_2 from Eq. (1). The labels are purely based on the quality of exposures when only E_1 is involved. In Fig. 4(c), the persons in the second image can be excluded if dynamic detection is enabled.

3. EXPERIMENTS

We assemble a comprehensive dataset of 120 groups multi-exposure image sequence from previous publications, Internet and our own capture. More results of our method are shown in the supplementary file. In this section, for the comparison, several typical MEF methods are selected [2, 5, 6, 7] (we collect the code from the authors and generate their results with same inputs). Both visual comparisons and objective evaluation are conducted in the experiments.

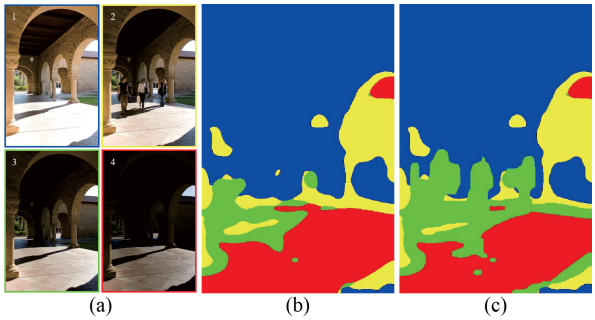
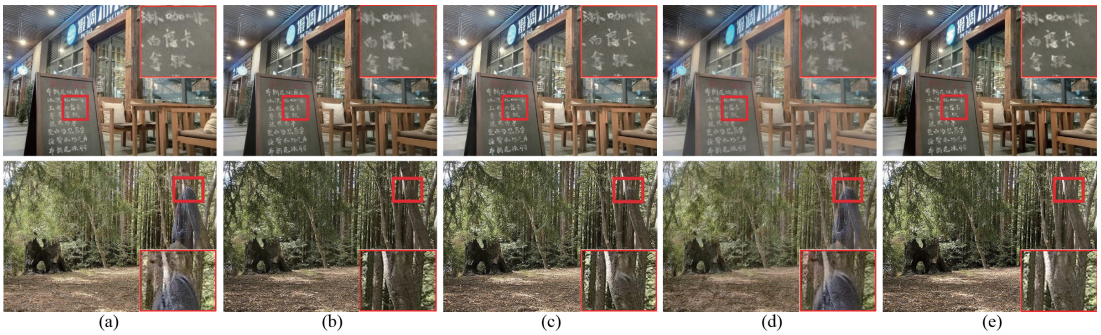


Fig. 4. (a) Input images where the third image is the reference. (b) Labels without dynamic term. (c) Final labels with dynamic term. Notably, if we want to keep the persons, we can select the second image as the reference.

3.1. Visual comparisons

We conduct two visual comparisons of our method with [2, 5, 6, 7]. The comparison results are shown in Fig. 5. The top row displays the results of a roughly aligned scene. There are some slight misalignments in nether parts of inputs. We found that other methods ((a)-(d) in top row) are not sensitive to such cases. They have different degree of fuzziness in the 'text' regions. Our method abandons generating results from every input. We select regions from single image which avoid blur artifacts effectively. The bottom row shows the results of a dynamic scene where exists a moving man. [2] and [7] have severe ghosting. [5, 6] can solve dynamic objects to some extent, while [6] still exists slight ghosting and [5] blurs the leaves in left region of their result. Our method does not involve synthesis process. It is likely to select areas continuously so it generally owns higher visual quality.



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

Fig. 5. Comparison with [2, 5, 6, 7]. (a) Results of [2]. (b) Results of [5]. (c) Results of [6]. (d) Results of [7]. (e) Our results. Please zoom in for a clearer observation.

Table 1. Image performance of different methods. The corresponding images are shown in supplement file.

	Index	Image1	Image2	Image3	Image4	Image5	Image6	Image7	Image8	Image9
[2]	Q_{MI}	0.489	0.262	**0.505**	0.411	0.446	**0.565**	0.874	0.57	0.408
	Q_{NCIE}	0.629	0.609	**0.632**	0.625	0.623	**0.635**	0.658	0.637	0.625
[5]	Q_{MI}	0.22	0.219	0.406	0.47	0.449	0.388	**0.875**	0.636	**0.462**
	Q_{NCIE}	0.623	0.608	0.629	0.627	0.624	0.628	**0.659**	0.641	**0.628**
[6]	Q_{MI}	0.314	0.222	0.287	0.321	0.361	0.326	0.842	0.567	0.277
	Q_{NCIE}	0.624	0.609	0.626	0.623	0.621	0.627	0.658	0.638	0.622
[7]	Q_{MI}	0.28	0.146	0.431	0.235	0.235	0.293	0.623	0.38	0.313
	Q_{NCIE}	0.623	0.607	0.63	0.618	0.619	0.626	0.649	0.631	0.622
Ours	Q_{MI}	**0.563**	**0.297**	0.47	**0.527**	**0.472**	0.441	0.861	**0.655**	0.446
	Q_{NCIE}	**0.635**	**0.612**	**0.632**	**0.632**	**0.626**	0.630	**0.659**	**0.643**	**0.628**

3.2. Objective assessments

We adopt two popular image fusion metrics Q_{MI} [28] and Q_{NCIE} [29] to evaluate the performances objectively. The metric Q_{MI} is defined as:

$$Q_{MI} = \frac{MI(A,F)}{H(A)+H(F)} + \frac{MI(B,F)}{H(B)+H(F)} \tag{8}$$

where A, B are inputs; F is fusion result; H represents the marginal entropy of an image; MI is mutual information between two images. Q_{MI} measures how well the original information from source image is preserved in the fused image. The large value of Q_{MI} indicates better results.

The nonlinear correlation entropy Q_{NCIE}, used as a nonlinear correlation measure, is defined as:

$$Q_{NCIE} = 1 + \sum_{i=1}^{K} \frac{\lambda_i}{K} \log_b \frac{\lambda_i}{K} \tag{9}$$

where b is determined by the intensity level; $\lambda_i (i = 1, ..., K)$, is the eigenvalues of the nonlinear correlation matrix. NCIE owns strong suitability as a measure for the nonlinear type of correlation of multiple variables. Similar to Q_{MI}, large Q_{NCIE} value indicates better results. Table 1 summaries the Q_{MI} and Q_{NCIE} values with respect to 9 examples. Our method can produce superior results on most of the cases.

In the comparison, we notice that if the original images exist large shaking and possess some depth variations, they cannot be aligned perfectly. Our method performs well with respect to these misaligned regions compared with other approaches. Because our seams can stitch different parts of the input images, which bypass the misaligned regions and hide them behind. Moreover, the seams can also bypass the dynamic objects, which naturally avoid the ghosting artifacts. By combining different parts of the input images, our method generate good results from roughly aligned images, thus relaxing the challenging alignment problem.

4. CONCLUSION

We have presented a method for accurately fusing multi-exposure images. We do not require high quality registration. We select good exposed regions from the roughly aligned images. Good seams are found to hide the misalignments when solving Poisson equation. The results are evaluated qualitatively and quantitatively to demonstrate its effectiveness.

5. ACKNOWLEDGE

This work has been supported by National Natural Science Foundation of China (61502079 and 61720106004).

References

[1] P. J. Burt and R. J. Kolczynski, "Enhanced image capture through fusion," in *Proc. ICCV*, pp. 173–182, 1993.

[2] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion," *Computer Graphics Forum.*, vol. 28, no. 1, pp. 382–390, 2007.

[3] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Conference on Computer Graphics and Interactive Techniques*, pp. 369–378, 1997.

[4] E. Reinhard, G. Ward, S. Pattanaik, and P. Debevec, *High dynamic range imaging : acquisition, display, and image-based lighting*. Princeton University Press, 2010.

[5] S. Li and X. Kang, "Fast multi-exposure image fusion with median filter and recursive filter," *IEEE Trans. on Consumer Electronics.*, vol. 58, no. 2, pp. 626–632, 2012.

[6] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. on Image Processing.*, vol. 22, no. 7, pp. 2864–2875, 2013.

[7] S. Paul, I. S. Sevcenco, and P. Agathoklis, "Multi-exposure and multi-focus image fusion in gradient domain," *Journal of Circuits, Systems and Computers.*, vol. 25, no. 10, p. 1650123, 2016.

[8] P. J. Burt and E. H. Adelson, "The laplacian pyramid as a compact image code," *IEEE Trans. on Communctions*, vol. 31, no. 4, pp. 532–540, 1983.

[9] O. T. Tursun, A. Erdem, and E. Erdem, "The state of the art in hdr deghosting: A survey and evaluation," *Computer Graphics Forum.*, vol. 34, no. 2, pp. 683–707, 2015.

[10] M. Granados, K. I. Kim, J. Tompkin, and C. Theobalt, "Automatic noise modeling for ghost-free hdr reconstruction," *Acm Transactions on Graphics*, vol. 32, no. 6, pp. 1–10, 2013.

[11] T. Jinno and M. Okuda, "Motion blur free hdr image acquisition using multiple exposures," in *Proc. ICIP*, pp. 1304–1307, 2008.

[12] K. Ma, H. Li, H. Yong, Z. Wang, D. Meng, and L. Zhang, "Robust multi-exposure image fusion: A structural patch decomposition approach," *IEEE Trans. on Image Processing.*, vol. 26, no. 5, pp. 2519–2532, 2017.

[13] J. Hu, O. Gallo, K. Pulli, and X. Sun, "Hdr deghosting: How to deal with saturation?," in *Proc. CVPR*, pp. 1163–1170, 2013.

[14] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman, "Robust patch-based hdr reconstruction of dynamic scenes.," *ACM Trans. Graphics.*, vol. 31, no. 6, pp. 1–11, 2012.

[15] Z. Cui, O. Wang, P. Tan, and J. Wang, "Time slice video synthesis by robust video alignment," *ACM Trans. Graphics.*, vol. 36, no. 4, pp. 1–10, 2017.

[16] K. Lin, N. Jiang, S. Liu, L. F. Cheong, M. Do, and J. Lu, "Direct photometric alignment by mesh deformation," in *Proc. CVPR*, pp. 2701–2709, 2017.

[17] F.-L. Zhang, X. Wu, H.-T. Zhang, J. Wang, and S.-M. Hu, "Robust background identification for dynamic video editing," *ACM Trans. Graphics.*, vol. 35, no. 6, p. 197, 2016.

[18] S. Liu, B. Xu, C. Deng, S. Zhu, B. Zeng, and M. Gabbouj, "A hybrid approach for near-range video stabilization," *IEEE Trans. Circ. Syst. Video Tech.*, vol. 27, pp. 1922–1933, 2016.

[19] H. Zimmer, A. Bruhn, and J. Weickert, "Freehand hdr imaging of moving scenes with simultaneous resolution enhancement," *Computer Graphics Forum.*, vol. 30, no. 2, pp. 405–414, 2011.

[20] N. K. Kalantari and R. Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes," *ACM Trans. Graphics.*, vol. 36, no. 4, pp. 1–12, 2017.

[21] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen, "Interactive digital photomontage," *ACM Trans. Graphics.*, vol. 23, no. 3, pp. 294–302, 2004.

[22] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proc. ECCV*, pp. 430–443, 2006.

[23] J. Shi and C. Tomasi, "Good features to track," in *Proc. CVPR*, pp. 593–600, 1994.

[24] H. Guo, S. Liu, T. He, S. Zhu, B. Zeng, and M. Gabbouj, "Joint video stitching and stabilization from moving cameras," *IEEE Trans. on Image Processing.*, vol. 25, no. 11, pp. 5491–5503, 2016.

[25] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, 2001.

[26] B. Zhang, Q. Liu, and T. Ikenaga, "Ghost-free high dynamic range imaging via moving objects detection and extension," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 459–462, 2015.

[27] Y. Li, J. Sun, C.-K. Tang, and H.-Y. Shum, "Lazy snapping," *ACM Trans. Graphics.*, vol. 23, no. 3, pp. 303–308, 2004.

[28] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," *Electronics letters.*, vol. 38, no. 7, pp. 313–315, 2002.

[29] Q. Wang, Y. Shen, and J. Jin, "Performance evaluation of image fusion techniques," *Image Fusion: Algorithms and Applications.*, vol. 19, pp. 469–492, 2008.