IEICE TRANS. INF. & SYST., VOL.E101–D, NO.2 FEBRUARY 2018

LETTER

End-to-End Exposure Fusion Using Convolutional Neural Network

Jinhua WANG[†,††a)], ***Member***, **Weiqiang WANG**[†], **Guangmei XU**[††], ***and*** **Hongzhe LIU**[†††], ***Nonmembers***

SUMMARY In this paper, we describe the direct learning of an end-to-end mapping between under-/over-exposed images and well-exposed images. The mapping is represented as a deep convolutional neural network (CNN) that takes multiple-exposure images as input and outputs a high-quality image. Our CNN has a lightweight structure, yet gives state-of-the-art fusion quality. Furthermore, we know that for a given pixel, the influence of the surrounding pixels gradually increases as the distance decreases. If the only pixels considered are those in the convolution kernel neighborhood, the final result will be affected. To overcome this problem, the size of the convolution kernel is often increased. However, this also increases the complexity of the network (too many parameters) and the training time. In this paper, we present a method in which a number of sub-images of the source image are obtained using the same CNN model, providing more neighborhood information for the convolution operation. Experimental results demonstrate that the proposed method achieves better performance in terms of both objective evaluation and visual quality.

key words: *exposure fusion, convolutional neural networks, fusion rule, activity level measurement*

1. Introduction

Digital cameras have a limited dynamic range that is lower than that in real-world environments. In high-dynamic-range (HDR) images, a bracketed exposure sequence can obtain the full dynamic range of the real scene. Fusing the multiple-exposure sequence into a high-quality image (called exposure fusion) has received considerable attention from the research community. There is another way to obtain the HDR scene information. This involves recovering a camera-specific response curve through multiple-exposure images and their exposure times, and obtaining a scene-related radiance map [1]. However, in most situations, the exposure time is unknown. Moreover, the intensities must be remapped to match the low dynamic range of the display device through a process called tone mapping [2]. In this paper, we focus on the first technique and propose a novel exposure fusion method that gives high-quality images of HDR scenes.

In the field of multi-exposure fusion, convolutional neural networks (CNNs) can be used to determine both the activity level measurement and fusion rule [3]. The framework shown in Fig. 1, is not an end-to-end fusion mechanism. The input is two image blocks of known size, and the network outputs two weight values of the corresponding blocks that determine the fusion process. To maintain more details of the scene, post-network segmentation and consistency verification steps are usually added, which greatly increases the complexity of the algorithm. Furthermore, the fusion framework can only fuse two images, and is not suitable for multi-exposure fusion processing. To solve these problems, this paper describes an end-to-end fusion strategy (EFCNN) in which the input is a sequence of multiple-exposure images. After passing through the convolution network, the fusion image is obtained directly, without the need for subsequent processing. The learnt mapping is represented as a deep CNN that takes multiple-exposure images as input and outputs a high-quality image. Our deep CNN has a lightweight structure, yet demonstrates high fusion quality. Furthermore, we know that for a given pixel, the influence of the surrounding pixels gradually increases as the distance decreases. If the only pixels considered are those within the convolution kernel neighborhood, the final result will be affected. To overcome this problem, previous methods have simply increased the size of the convolution kernel, which increases the complexity of the network (too many parameters) and the training time. This paper presents a novel technique whereby we obtain a number of sub-images of the sources using the same model, providing more neighborhood information for the convolution operation.

The remainder of this paper is organized as follows. Some related work is discussed in Sect. 2. Section 3 provides a detailed explanation of our proposed method. Exper-

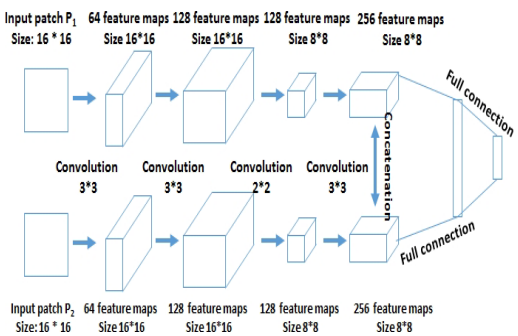

Fig. 1 Conventional fusion framework using deep convolution network.

Manuscript received August 8, 2017.
Manuscript revised October 20, 2017.
Manuscript publicized November 22, 2017.

[†]The authors are with School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China.
[††]The authors are with College of Information Technology, Beijing Union University, Beijing 100101, China.
[†††]The author is with Beijing Key Laboratory of Information Service Engineering, China.
a) E-mail: xxtwangjinhua@buu.edu.cn

DOI: 10.1587/transinf.2017EDL8173

imental results and performance evaluations are presented in Sect. 4. Section 5 concludes the paper.

2. Related Work

A number of multi-exposure fusion methods have been proposed. Mertens et al. [4] presented a method to fuse multiple-exposure images based on processing separate Laplacian pyramids in the R, G, and B channels. The results contain brightness changes that are not consistent with the original source images. These are caused by significant changes in brightness among images with different exposure times. Goshtasby [5] proposed an exposure fusion method from multiple-exposure images of a static scene. His approach blends the image blocks from a specific domain by selecting uniform image blocks that contain the most useful information. Because a block may span different objects, this approach cannot handle object boundaries. Gu et al. [6] proposed a gradient field multi-exposure image fusion method for HDR image visualization. The advantage of this method is its computational efficiency and robustness. Only two parameters are used, and they can generally be set to default values. However, the metric to measure the distance between intensities should be improved to reduce the need for tedious gradient modification. Song et al. [7] synthesized an exposure fusion image using a probabilistic model that preserves the luminance levels and suppresses reversals in the image luminance gradients. Shen et al. [8] proposed a novel hybrid exposure weight measurement that is guided not only by a single image's exposure level, but also by the relative exposure level among different exposure images using a boosting Laplacian pyramid.

3. Convolutional Neural Networks for Exposure Fusion

In the proposed method, we use the CNN to achieve end-to-end exposure fusion. "End-to-end" means that, through the network operation, a fused image is directly generated. Given the multiple-exposure fusion images denoted as Y, our goal is to recover an image $F(Y)$ that is as similar as possible to the ground truth image X. We use the CNN to learn the mapping F that denotes the relationship between under-/over-exposed images and the standard-exposure image. An overview of the network is depicted in Fig. 2.

We first describe the process of the framework (denoted by the black line in Fig. 2): 1) Obtain gray images of the input multiple-exposure images and denote them as Y. From Fig. 2 we can see that the gray images of the source images are served as input for the network, and the output of the network is the gray image. How to obtain the color resulting image, the traditional methods usually deal with the R, G, B channel separately. In our method, we use the strategy in our former work presented in Ref. [10] to obtain the color fusion resulting image. 2) The first convolution layer is expressed as an operation F_1:

$$F_1(\mathrm{Y}) = \max(0, W_1 * \mathrm{Y} + B_1) \tag{1}$$

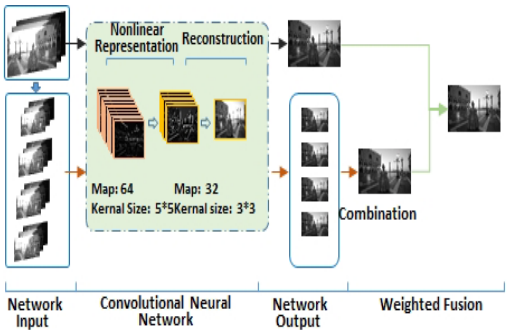

Fig. 2 Our framework of the end-to-end exposure fusion method.

where W_1 and B_1 denote the filters and biases, respectively. W_1 is an $f_1 \times f_1$ convolution kernel filter with n_1 elements, where f_1 is the spatial size of the filter. We can see that the first convolution layer uses W_1 to apply n_1 convolutions to the image, and each convolution has a kernel size of $f_1 \times f_1$. The output is composed of n_1 feature maps. B_1 is an n_1-dimensional vector in which each element is the bias of the corresponding filter. The commonly used Rectified Linear Units (ReLU) are adopted for the filter responses, making the convergence much faster while still presenting good quality. 3) The first layer extracts an n_1-dimensional feature. In the second operation, we map each of these n_1-dimensional features into an n_2-dimensional feature. The formula is defined as:

$$F_2(\mathrm{Y}) = \max(0, W_2 * F_1(\mathrm{Y}) + B_2) \tag{2}$$

where W_2 is an $f_2 \times f_2$ convolution kernel filter containing $n_1 \times n_2$ elements. B_2 is an n_2-dimensional vector in which each element is the bias corresponding to a filter in this layer. 4) The last operation occurs in the reconstructive layer. In traditional methods, the final image is often averaged by the n_2 feature maps. The averaging can be considered as a predefined filter on a set of feature maps, that is, the weight is same for all inputs. This may lead to a "flattening" of the final fused image. In our method, we define a convolution layer to generate the final fused image. The formula is defined as:

$$F(\mathrm{Y}) = W_3 * F_2(\mathrm{Y}) + B_3 \tag{3}$$

where W_3 is $f_3 \times f_3$ convolution kernel filter with n_2 number.

To learn the end-to-end mapping function F, the parameters $\Theta = \{W_1, W_2, W_3, B_1, B_2, B_3\}$ must be estimated. In our method, this is achieved by minimizing the loss between the fused images $F(\mathrm{Y}; \Theta)$ and the corresponding ground truth images X (standard exposure illumination). Given a set of training images, we use the Mean Squared Error (MSE) as the loss function:

$$L(\Theta) = \frac{1}{n} \sum_{i=1}^{n} \|F(\mathrm{Y}; \Theta) - \mathrm{X}_i\|^2 \tag{4}$$

where n denotes the number of training samples.

To examine the performance of our method, training examples are generated from the images in the ILSVRC

2012 validation image set, which contains 50,000 high-quality natural images. To obtain under-exposed images from this image set, we select random numbers between 0.4 and 1 to rescale the pixel intensity, and multiply the natural image by the number. In the same way, we select random numbers between 1.2 and 1.8, and multiply the natural image by this number to obtain over-exposed images. For each under-exposed/over-exposed image and the natural image, we randomly sample pairs of 33×33 patches. Thus, we obtain a total of 744,175 pairs of patches from the training set.

The loss function defined in Eq. (4) is minimized using stochastic gradient descent. In our training procedure, the batch size is set to 128 and the weight decay and momentum are set to 0.0005 and 0.9, respectively. The weights are updated using the rule:

$$\vartheta_{i+1} = 0.9 \cdot \vartheta_i - 0.0005 \cdot \alpha \cdot w_i - \alpha \cdot \frac{\partial L}{\partial w_i}, w_{i+1} = w_i + \vartheta_{i+1} \quad (5)$$

where ϑ is the momentum variable, i is the iteration number, α is the learning rate, L is the loss function, and $\partial L/\partial w_i$ denotes the derivative of the loss with respect to the weight at w_i. We train the model using the popular deep learning framework Caffe [9]. The learning rate is equal for all layers and is initialized as 0.0001 as Ref. [3].

According to the design of the network described above, the process of the black arrow in Fig. 2 shows that, with different exposures of the image sequence through the network, we can reconstruct a fusion image. However, in the convolution process, each pixel is computed using the pixels within the convolution kernel. To simplify the description, we use the top part in Fig. 3 to denote an image of size 16×16. For a given pixel, shown in the red box, if we use a convolution kernel of size 3×3, only the pixels on the side of the black box are calculated in the convolution process. For a given pixel, the influence of the surrounding pixels gradually increases as the distance decreases. If only the pixels in the convolution kernel neighborhood are considered, the final result will be affected. To overcome this problem, previous methods increase the size of the convolution kernel, but this increases the complexity of the network (too many parameters) and the training time. In the proposed method, we first sample the original image to obtain a number of sub-images, thus providing more neighborhood information for the convolution operation. For example, we down-sample the original image of size N, and obtain N^2 sub-images. To simplify the description, take $N = 2$ as an example. As shown in the bottom part of Fig. 3, we have the four sub-images denoted by different colors. For the same pixel shown in the red box in the yellow sub-image, the convolution operation is applied to pixels 1-8 in the original image. This increases the influence of the neighboring pixels corresponding to the inside of the box in the original image. As N increases, the influence of neighborhood pixels can be further enhanced, and the design of the network will not be affected. As shown in Fig. 2, we obtain four fused sub-images, and combine these to give the fused image. The fused image is obtained by convolution using neighborhood pixels that are far away (with a tentative weight of 0.3) and the fused image using the original sequence (with a tentative weight of 0.7). Using the two fused images, we obtain the final resulting image.

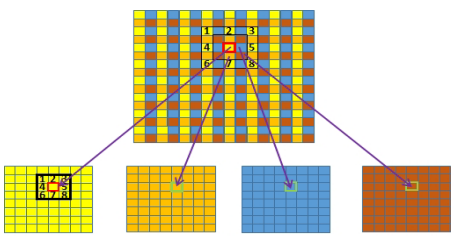

Fig. 3 Method of extending the neighborhood.

4. Experiments

Three objective criteria were used to quantitatively evaluate the performance of the exposure fusion methods. The first criterion is mutual information (MI), defined as the sum of mutual information between each input image and the fused image. The second criterion is $Q^{A,B,\ldots,Z/F}$, which measures the amount of edge information transferred from the source images to the fused image. The third criterion is entropy, which measures the overall information in the fused image. Reference [10] provide more details on these criteria.

We use the "grandcanal" source exposure images to verify the effectiveness of EFCNN. A visual comparison is shown in Fig. 4. The top row shows the source images, image (B) is that obtained by the method in [4], image (C) is that obtained by the method in [5], and image (D) is that obtained by the proposed EFCNN with sampling. We can see that there is a brightness change in images (B) and (C), obtained by the methods of [4] and [5], respectively. Although more cloud details can be seen, significant chrominance information has been lost. As a whole, images (B) and (C) look a little unnatural to the human eye. Image (D), obtained by EFCNN, provides a warm perception, although some details of the clouds in the sky have been lost.

To demonstrate the effectiveness of the EFCNN sampling strategy in preserving more details of the source im-

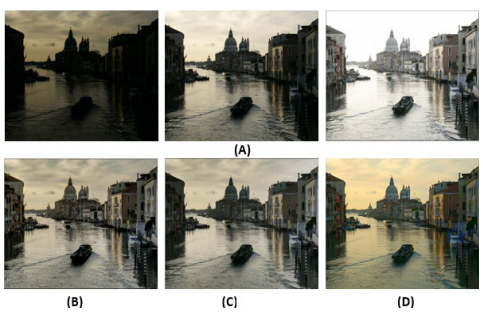

Fig. 4 Comparison with typical exposure fusion methods. (A) Source images, (B) Mertens et al. [4], (C) Goshtasby [5], (D) EFCNN with sampling. Images courtesy of HDRsoft.

Fig. 5 Source images to verify the sampling strategy of the EFCNN.

Table 1 Comparison results for $Q^{A,B,...,Z/F}$, MI, and *Entropy* for validating the effectiveness of sampling in EFCNN.

Image	Method	Q	MI	*Entropy*
Fig. 5	Without_ sampl_EFCNN	0.4639	10.6545	7.3603
Fig. 5	With_ sampl_EFCNN	0.4876	11.2966	7.4078

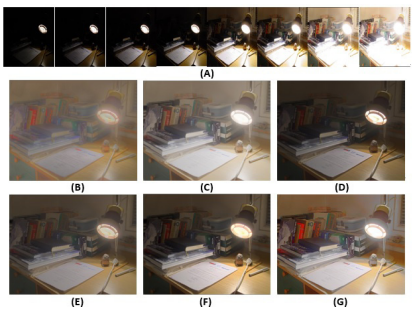

Fig. 6 Comparison results with tone-mapping operators. (A) Source images, (B) Fattal, (C) Drago, (D) Krawczyk, (E) Ashikhmin, (F) Reinhard, (G) EFCNN with sampling. Images courtesy of Cadik et al. [2].

ages, the $Q^{A,B,...,Z/F}$, MI, and *Entropy* criteria using five source images in Fig. 5 are computed for the fused results obtained by EFCNN with sampling and without sampling. The statistical quantitative results are presented in Table 1. It is clear that the EFCNN with sampling provides better fusion performance in terms of the quantitative criteria. The $Q^{A,B,...,Z/F}$ value of EFCNN with sampling is 0.4876, whereas that without sampling is 0.4639. The MI value of EFCNN with sampling is 11.2966, compared with 10.6545 without sampling. The *Entropy* value of EFCNN with sampling is 7.4078, which is again better than the value of 7.3603 achieved by EFCNN without sampling. Based on the above analysis, we can see that the proposed sampling strategy is effective for preserving more details.

Similar to the proposed method, the aim of tone mapping is to acquire high-quality images for display on ordinary devices. In Fig. 5, the results of our proposed EFCNN are compared with the output from some typical tone-mapping methods. The results obtained by Fattal (B), Drago (C), Krawczyk (D), Ashikhmin (E), and Reinhard (F), as summarized by Cadik et al. [2], are used as representative tone-mapping methods. We can see from Fig. 6 that image (B) has lost some local contrast, resulting in an image that looks unnatural to the human eye. Image (C) is washed out and some chrominance information has been lost. In image (D), certain details of the books behind the lamp in the background are not visible. Reinhard et al.'s method (F) generates quite similar results to image (E), obtained by Ashikhmin, in terms of contrast and detail preservation. We can see that some details in images (E) and (F) have been lost compared to image (G), obtained by EFCNN with sampling. We also provide a quantitative comparison of the images in Fig. 5 to demonstrate the performance of EFCNN with sampling. However, criteria $Q^{A,B,...,Z/F}$ and MI are input-related, and the inputs for exposure fusion and tone mapping are different. Thus, we use the *Entropy* criterion, which is not related to the input, to validate the effectiveness of our method. The resulting values for *Entropy* are presented in Table 2. From this table, we can see that EFCNN is competitive with other typical tone-mapping methods in terms of detail preservation.

Table 2 Entropy comparison between EFCNN and tone mapping methods.

Image	(B)	(C)	(D)	(E)	(F)	(G)
Entropy	6.3108	7.3101	6.6746	6.8799	7.3115	7.4755

5. Conclusion

In this paper, we have described the use of CNNs to learn an end-to-end exposure fusion model. The image feature representation and fusion rules are obtained simultaneously using the learning approach. To preserve more details, we have proposed a technique that uses more neighborhood pixels to calculate the convolution without changing the network structure. Experiments show that the method is comparable or even better than existing exposure fusion methods.

Acknowledgements

This work was supported by National Nature Science Foundation of China (No. 91420202, No. 61572077, No. 61372148).

References

[1] P.E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," Proc. SIGGRAPH '97, pp.369–378, 1997.
[2] M. Čadik, M. Wimmer, L. Neumann, A. Artusi, "Evaluation of HDR tone mapping methods using essential perceptual attributes," Computers & Graphics, vol.32, no.3, pp.330–349, 2008.
[3] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," Information Fusion, vol.36, pp.191–207, 2017.
[4] T. Mertens, J. Kautz, and F.V. Reeth, "Exposure fusion," Proc. Pacific Graphics, pp.382–390, Maui, Hawaii, 2007.
[5] A.A. Goshtasby, "Fusion of multi-exposure images," Image and Vision Computing, vol.23, no.6, pp.611–618, 2005.
[6] B. Gu, W. Li, J. Wong, M. Zhu, and M. Wang, "Gradient field multi-exposure images fusion for high dynamic range image visualization," J. Visual Communication and Image Representation, vol.23, no.4, pp.604–610, 2012.
[7] M. Song, D. Tao, C. Chen, J. Bu, J. Luo, and C. Zhang, "Probabilistic exposure fusion," IEEE Trans. Image Proces., vol.21, no.1, pp.341–357, 2012.
[8] J. Shen, Y. Zhao, S. Yan, and X. Li, "Exposure fusion using boosting using Laplacian pyramid," IEEE Trans. Cybern., vol.44, no.9, pp.1579–1590, 2014.
[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," ACM Multimedia, pp.675–678, 2014.
[10] J. Wang, H. Liu, and N. He, "Exposure fusion based on sparse representation using approximate K-SVD," Neurocomput., vol.135, pp.145–154, 2014.