MULTI-EXPOSURE IMAGE FUSION BASED ON STRUCTRUE CONSISTENSY

[1] *Yen-Kai, Fan* (范延愷), [1] *Chiou-Shann, Fuh* (傅楸善)

[1] Department of Computer Science and Information Engineering,
National Taiwan University, Taipei Taiwan,
E-mail: B04902078@ntu.edu.tw fuh@csie.ntu.edu.tw

ABSTRACT

Multi-Exposure image Fusion (MEF) is a widely used technique to enhance the quality of images with different exposure times by fusing them. In recent years, some people proposed different MEF algorithms, and other people devoted themselves into evaluating the quality of images generated by those algorithms, that is, they proposed some Image Quality Assessment (IQA) algorithms for MEF images. The IQA model proposed by Kede Ma [1] is based on some considerable agreement among human subjects on the quality of MEF images. Inspired by their observation and method., we would like to propose a novel weighting function which is also based on patch structure consistency.

Keywords: *Multi-exposure image fusion (MEF)*

1. INTRODUCTION

Recently, Multi-Exposure image Fusion (MEF) becomes a widely adopted quality enhancement technique. It takes a sequence of images with different exposure times as input, generates a sequence of weight maps, fuses the input image sequence by the weight maps, and outputs an image which is more informative than any of the images in the input sequence.

The dynamic range of natural scene is much greater than our commercial products. That is, photographs cannot present the world just as human eyes. To overcome this problem and make photographs closer to what we see in our daily life, High Dynamic Range (HDR) imaging technique is proposed. HDR technique also takes a sequence of images with different exposure times as input and the most popular method is reconstructing an HDR image by estimating the response curve [2]. The next step of this technique tone-mapping, in this step, the HDR image is remapped into a Low Dynamic Range (LDR) image for display, since most of our devices are not able to display an image with high dynamic range. There are some different ways for implementation, Local operator spatially variously remaps the intensities to compress the dynamic range. This often generates more pleasing result, sometimes will also make it looks unnatural. Global operator spatially uniformly remaps the intensities to compress. It is faster than using local one, however, it produces unpleasing images sometimes [3-9].

Different from typical HDR imaging technique which computes an HDR image first and remaps it into LDR image for display, MEF method skips the step of computing a HDR image, immediately fused the images into a high-quality, LDR image. There are some advantages of this kind of methods against traditional HDR technique, for example, it simplifies the pipeline and also allows users add some special effects by adding a flash image [10].

After the first MEF algorithm was released, many papers about novel algorithms came out [11-15], some of them focused on how to generate reasonable weight maps, another focused on how to make the fused images without unpleased effects. They also find out some problems about this topic. However, even many algorithms are proposed, measuring the quality of the results of each MEF algorithm is still a big problem since there is not a ground truth good enough to compare with. In [1], they proposed an IQA model to measure the quality of MEF images based on their subjective user study and adopted contrast and structure consistency as their parameters. These parameters make their IQA model reliable.

With a reliable IQA model for MEF, the quality of fused images produced by different MEF algorithms can be evaluate fairly. Moreover, the parameters they adopted are used to measure the quality of patches in images. That is, the parameters can also be adopted to be our quality measures to generate the weight maps of MEF. Thus, we design a novel MEF algorithm based on structure consistency. We also adopt contrast and the well-exposedness as the quality measures to generate weight maps.

2. RELATED WORK

2.1. Multi-Exposure Image Fusion

Fig. 1: Overview of Exposure Fusion. The first row is an input image sequence. The second row is the corresponding weight maps reflect the image quality by contrast, structure similarity, and luminance. Image courtesy is [10]

MEF problem can be formulated:

$$R_{ij} = \sum_{k=1}^{N} W_{ij,k} I_{ij,k} \qquad (1)$$

where N is the number of images in a input sequence; R is the fused image; I and W are the input image sequence and the corresponding weight map sequence. The subscripts i, j, k refers to pixel (i, j) in the k-th image.

Based on (1), we can reconstruct the sequence to a fused image. Unfortunately, there will be some abrupt effects and halos in the fused image if we fuse the image sequence straight forward. One of the most successful strategies to solve this problem is Laplacian pyramid decomposition based method introduced by Burt and Adelson [6]. Not only exposure fusion, this method can be employed by many different applications. Multi-Scale image Fusion (MSF) methods based on it also became popular result from its effectiveness. After Mertens et al [10] proposed their algorithm which is effective and easy to understand, many algorithms adopting MSF methods based on different theories are proposed to improve the quality of fused image [11, 12, 13, 15]. Besides, some people proposed Single-Scale image Fusion (SSF) methods to improve the performance of algorithm [14].

2.2. Image Database

The database contains 14 natural image sequences. All of the sequences contain at least 3 images from

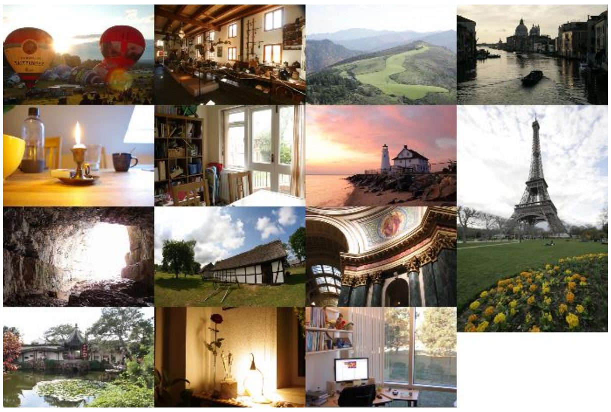


Fig. 2: Image sequences in the database. Each sequence is represented by the image which is close to normal exposure time.

Table 1: Details of the sequences

Source	N	Resolution (pixels)	Image Courtesy
Balloons	9	512*339*3	Erik Reinhard
Candle	10	512*364*3	HDR projects
Cave	4	512*384*3	Bartlomiej Okonek
Chinese garden	3	512*340*3	Bartlomiej Okonek
Farm house	3	512*341*3	HDR projects
House	4	512*340*3	Tom Merten
Kluki	3	512*341*3	Bartlomie Okonek
Lamp	6	512*342*3	HDR projects
Landscape	3	512*341*3	HDRsoft
Light house	3	512*340*3	HDRsoft
Madison	30	512*384*3	Chaman Singh Verma
Tower	16	341*512*3	Jacques Joffre
Office	6	512*340*3	MATLAB
Venice	3	512*341*3	HDRsoft

underexposed to overexposed images. In addition,

natural scenery, artificial building, indoor and outdoor view are all included in the database. For example, one of the sequences is shown in the first row in Fig. 1. Details including the size, resolution, and number of the images in each sequence is shown in Table 1. For each sequence, the image close to normal exposure time is chosen and shown in Fig. 2.

3. METHOD

3.1. Quality Measures

There are two points should be mentioned at first. One is that we use $\{x_{ij,k}\} = \{x_{ij,k} | 1 \leqq k \leqq N\}$ to denote the set of image patches in the same location (i, j) of N different input images. The other is the color space we adopt is Lab color space since it is known as closer to human visual system than other color spaces.
In addition, different from algorithms using RGB space [10], we only generate one weight map for each image instead of one weight map for each channel since L channel is the only channel related to exposedness. Due to this property, comparing with those algorithms using RGB color space and deal the color channels separately, using Lab color space also make our algorithm faster.
Learning from Structure Similarity (SSIM) approach and [1], we can decompose an image patch into three components, contrast (denoted by c), structure (denoted by s), and luminance (denoted by l). Their mathematical definition is denoted as follow.

$$\begin{aligned} x_{ij,k} &= \| x_{ij,k} - \mu_{x_{ij,k}} \| \cdot \frac{x_{ij,k} - \mu_{x_{ij,k}}}{\| x_{ij,k} - \mu_{x_{ij,k}} \|} + \mu_{x_{ij,k}} \\ &= \| \tilde{x}_{ij,k} \| \cdot \frac{\tilde{x}_{ij,k}}{\| \tilde{x}_{ij,k} \|} + \mu_{x_{ij,k}} \\ &= c_{ij,k} \cdot s_{ij,k} + l_{ij,k} \end{aligned} \qquad (2)$$

where $\mu_{x_{ij,k}}$ is the mean value of the patch, $\tilde{x}_{ij,k}$ is the mean-removed version of $x_{ij,k}$. Contrast and luminance are scalars and they are represented by l^2 norm and mean intensity of $\tilde{x}_{ij,k}$. Structure is an unit-length vector $s_{ij,k} = \tilde{x}_{ij,k} / \| \tilde{x}_{ij,k} \|$.
We also adopt these three components as the quality measure used to measure the quality of each patch.

3.1.1. Contrast

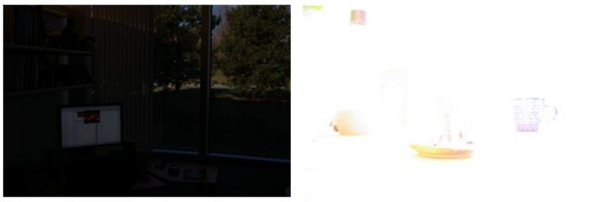
The visibility of a patch depends on its contrast. Also, the patch becomes more visible with larger contrast. For MEF problems, the inputs are all unprocessed photographs. Thus, there will not be any unrealistic local structure with high contrast to confuse us. If a patch has larger contrast than the other, it also contains more detail. Due to this property, the value of $c_{ij,k}$ can be used straight forward.

$$C(x_{ij,k}) = c_{ij,k} \qquad (3)$$

3.1.2. Structure

The structure of a local patch is denoted by a unit-length vector $s_{ij,k}$, where $1 \leqq k \leqq N$. However, the length of patch vector is also an important parameter to evaluate the weight, so we prefer using $\tilde{x}_{ij,k}$, the zero mean form, to evaluate the contribution of each patch of input images at this position.

First, if the structure vectors of the N patches are much different from each other, they should have similar eight.



(a) (b)

Fig. 3: (a) Under-exposed image in sequence "Office". (b) Over-exposed image in sequence "Candle".

On the other hand, if the structure vectors are similar, the patch with longer $\tilde{x}$ should contribute more to make the result contain more details. Thus, we employ a power weighting function, where exponent p represents the structure similarity:

$$S(x_{ij,k}) = \| \tilde{x}_{ij,k} \|^p \qquad (4)$$

To measure p, we compute

$$R(\{\tilde{x}_{ij,k}\}) = \frac{\| \sum_{k=1}^{N} \tilde{x}_{ij,k} \|}{\sum_{k=1}^{N} \| \tilde{x}_{ij,k} \|} \text{ and } p = \tan \frac{\pi R}{2} \qquad (5)$$

where $R(\{\tilde{x}_{ij,k}\}) \in [0, 1]$. If the structure vectors are similar, in the extreme case $R = 1$, which means the direction of structure vectors are all the same, p will be ∞. Practically, since overflow problem will be encountered in the next step, we let $p = 0.95$ in this situation. In the other extreme case that $R = 0$, which means they have no consistency, p will be 0. That is, the weight exponent p of each input at this pixel is all the same.

3.1.3. Luminance (Well-exposedness)

First, to evaluate the luminance of images, we adopt Lab color space instead of RGB color space since it is known as closer to human visual system.
Luminance reveals the object color and how well the pixel is exposed. Mertens et al. [10] used Gauss curve to evaluate how close 0.5 and the pixel value on each color channel separately to measure if a pixel is well-exposed or not. That is, the function they used for evaluating well-exposedness is as follow, where σ equals to 0.2 in their implementation.

$$L_{Mertens}(I_{ij,k}) = e^{\frac{-(I_{ij,k} - 0.5)^2}{2\sigma^2}} \qquad (6)$$

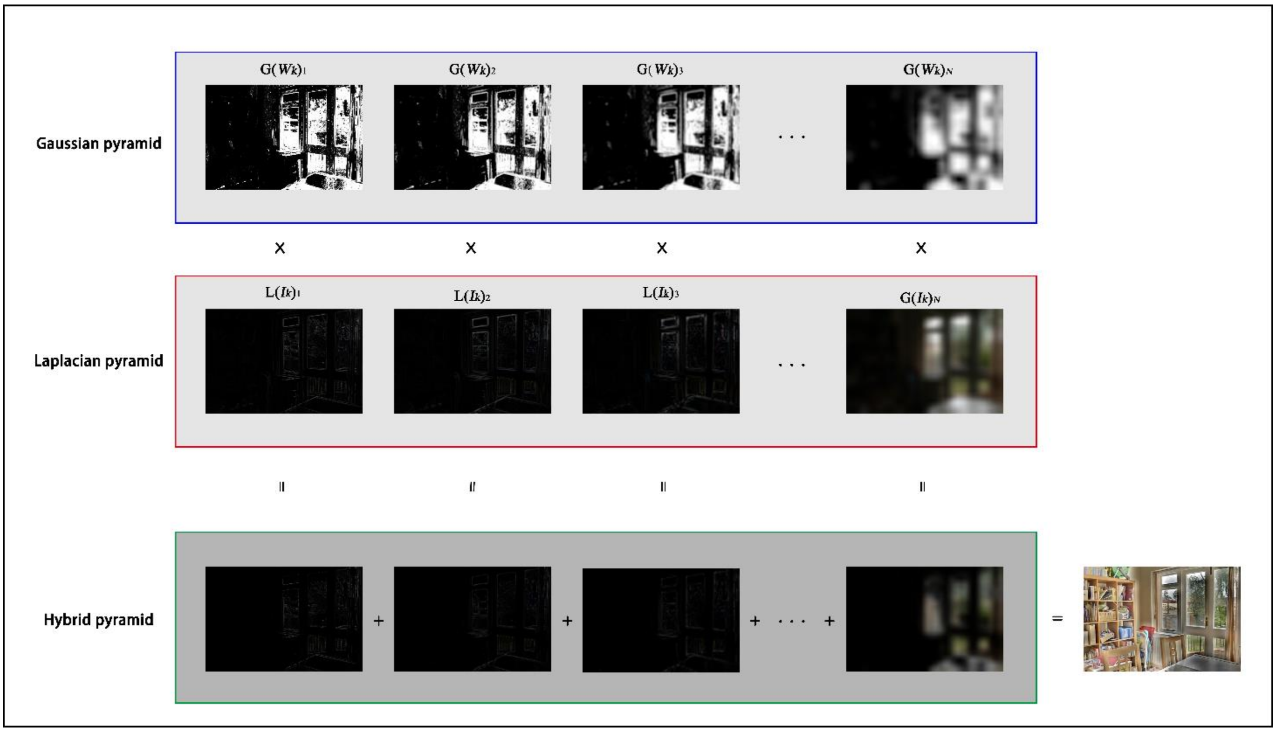


Fig. 4: Fusion pipeline. Each level of Gaussian and Laplacian pyramid have been resized to the original size of the image.

Nevertheless, Q. Wang [15] found out there is a problem in $L_{Mertens}\left(I_{ij,k}\right)$ that sometimes the lightest pixel in the under-exposed image is still much lower than 128. Therefore, the darkest pixel in the over-exposed image may be much greater than 128. In our case, since we adopt Lab color space, the middle intensity value is 50 instead of 128.

For example, the brightest pixel value in the under-exposed image in sequence “Office” is 36, and the darkest value in the over-exposed image in sequence “Candle” is 73 (Fig. 3).

This problem results in lower contrast in the fused image. Thus, we change $L_{Mertens}\left(I_{ij,k}\right)$ to $L\left(x_{ij,k}\right)$, a function of $x_{ij,k}$, and replace the mean value of Gauss curve by $M(\mu_k)$.

$$L\left(x_{ij,k}\right) = e^{\frac{-\left(l_{ij,k} - M(\mu_k)\right)^2}{2\sigma^2}} \tag{7}$$

$$\begin{aligned} M(\mu_k) &= 50 + t * (\mu_k - 50) \\ &= (1-t) * 128 + t * \mu_k \\ &\quad t \in [0,1] \end{aligned} \tag{8}$$

where μ_k is the mean intensity of the k-th image, t is a constant between zero and one, and σ equals to 40 in our implementation. For an extreme case which $t = 0$, $M(\mu_k)$ will equals to 50, the only difference between $L_{Mertens}\left(I_{ij,k}\right)$ and $L\left(x_{ij,k}\right)$ is that pixel intensity $I_{ij,k}$ is replaced by patch mean intensity $l_{ij,k}$. For the other extreme case which $t = 1$, $M(\mu_k)$ will equals to μ_k. In theory, the result does not meet our expectation obviously since for the problem we have mentioned, the brightest pixel in the under-exposed image and the darkest pixel in the over-exposed image do not have the largest weight in the corresponding images. In implementation, we suggest that take t between 0.3 and 0.5.

After evaluating the input images by these three quality measures, the weight map can be constructed by multiplying them together:

$$W_{ij,k} = C\left(x_{ij,k}\right)^{\omega_C} \times S\left(x_{ij,k}\right)^{\omega_S} \times L\left(x_{ij,k}\right)^{\omega_L} \tag{9}$$

where C, S, L are three weighting component and $\omega_C, \omega_S, \omega_L$ are the weighting exponents for C, S, and L. By choosing different ω 's, we can control the contribution of weighting components in the final weight map. Practically, we take each $\omega = 1$, that is, three weighting components are as important as other.

To obtain a consistent result, we should normalize the value of $W_{ij,k}$ of patches in N images that they sum to one:

$$\widehat{W}_{ij,k} = \left[\sum_{l=1}^{N} W_{ij,l}\right]^{-1} W_{ij,k} \tag{10}$$

where $\widehat{W}$ is the normalized weight map.

3.2 Fusion

Fig. 5: Fused images from all the sequences in our database, where t = 0.5.

We also adopt MSF method based on Laplacian decomposition to fuse the images [10]. The only difference is that we adopt Lab color space instead of RGB color space. It can be expressed as:

$$R_{MSF} = \sum_{l=1}^{M+1} R_l \; and \; M = \min\{log_2 W, \; log_2 H\} \quad (11)$$

where R_l is the l-th level of Fused pyramid and M is the number of levels. R_l is defined as below:

$$R_l = \sum_{k=1}^{N} G\left(\widehat{W}_k\right)_l L(I_k)_l \quad (12)$$

where $G(\cdot)_l$ and $L(\cdot)_l$ denotes the l-th level of Gaussian pyramid and Laplacian pyramid, I_k and $\widehat{W}_k$ denote the k- th image and the corresponding normalized weight map. The fusion pipeline of Laplacian decomposition based method is shown in Fig. 4.

4. RESULTS

Expect for we mention, three quality measures of our results are equally weighted, that is, $\omega_C = \omega_S = \omega_L = 1$.

4.1 Quality

All of the fused images from our database is shown in Fig. 5. Compare with the results generated by proposed method and Mertens et al. [10], ours preserve more details and also have less over-exposed and under-exposed regions. These two properties are resulted from patch-based quality measurement and replacing the original function to $M(\mu_k)$.

Patch-based method deals the pixels locally, this property leads to larger contrast in high frequency areas such as those with textures and edges. Even though it also results in some abrupt transitions in the weight maps, for example, the weight maps of “House” shown

(a) (b)

Fig .6: The first row is fused image of “Tower” produced by (a) our technique (t = 0.5) and (b) Mertens et al. [10], other rows are parts of the fused images.

in Fig. 1, multi-scale fusion is able to solves this problem by pyramid-based method in the final fused image just as what has been mentioned in [14].

$M(\mu_k)$ replaces the mean value of Gauss curve and it is closer to the mean value of images. Contrast becomes larger since $M(\mu_k)$ gives larger weight to the bright areas in under-exposed images and dark areas in the over-exposed images.

In Fig. 6, we show the difference which we have mentioned above. In the second row, clouds in over-exposed area can be seen clearly and other area has larger contrast. In the third row, the structures of reinforcement bars are also able to be observed apparently.

Fig. 7: Fused image of "Balloon" with different *t*'s (0.1 ~ 1.0).

Table 2: Computation time of different sequences

Source	$h * w * N$	Init. (s)	Weighting (s)	Fusion (s)	Total (s)	Total/N (s)
Balloons	512*339*9	0.268592	0.421758	0.078955	0.769305	0.085478
Candle	512*364*10	0.164231	0.481740	0.092931	0.738902	0.073890
Cave	512*384*4	0.062130	0.207881	0.042976	0.312986	0.078247
Chinese garden	512*340*3	0.050748	0.139920	0.028985	0.219654	0.073218
Farm house	512*341*3	0.048008	0.140923	0.029962	0.218893	0.072964
House	512*340*4	0.061987	0.185895	0.037977	0.285853	0.071463
Kluki	512*341*3	0.044966	0.142920	0.033979	0.221866	0.073955
Lamp	512*342*6	0.091208	0.271846	0.053984	0.417038	0.069506
Landscape	512*341*3	0.044027	0.167904	0.029982	0.211914	0.070638
Light house	512*340*3	0.044993	0.137922	0.031979	0.214893	0.071631
Madison	512*384*30	0.534254	1.522251	0.293987	2.350492	0.07835
Tower	341*512*16	0.2533430	0.732598	0.148922	1.134863	0.070929
Office	512*340*6	0.102942	0.282897	0.059973	0.445811	0.074302
Venice	512*341*3	0.049179	0.144937	0.029963	0.224864	0.074955

However, how to choose a proper *t* is still a problem. In Fig. 7, we show ten results of "Balloon" produced by different *t*'s from 0.1 to 1.0. With a larger *t*, more details are preserved and the local contrast will become larger. However, the fused image will also become more unrealistic since the global contrast become smaller, which is much different from what we see in our everyday life. After our observation, expect for some special purposes, we suggest choose *t* from 0.3 to 0.5. In this range, the fused images are more informative and still close to our visual experience.

4.2 Performance

We have split our algorithm into three parts (initialization, weighting, and fusion), and measured the computation time of each parts. It is shown in Table 2, where initialization includes converting color space to Lab color space, measuring the mean luminance, and extracting patches; weighting includes constructing the weight map from three measurements; fusion includes Laplacian decomposition and re-converting color space to RGB color space for display.

Since the sizes of all images are almost same, we can easily observe that the total computation time is almost directly proportional to *N*.

We have also measured the computation time of sequences with different sizes and show it in Table 3.

Table 3: Computation time of different sizes. Image courtesy is “Farmhouse”.

$h * w * N$	Init. (s)	Weighting (s)	Fusion (s)	Total (s)
512*341*3	0.044993	0.137922	0.031979	0.214893
1024*682*3	0.173447	0.592661	0.146914	0.913023
2048*1364*3	0.658129	2.441949	0.662190	3.722267
4096*2728*3	2.523380	12.057740	2.955227	17.5636347

To compare computation time precisely, the larger images are resized from the original images. Even though our algorithm can run in real-time for small sizes, we expect our algorithm can run in real-time even on as mobile device with 4K images. We also observed that the quality of large image is lower than the original image since we use the same patch size for each of them. For large images, most of the patches look like low frequency area. This property makes two of the quality measurement, contrast and structure, become useless.

5. CONCLUSION AND FUTUREWORKS

We provide a patch-based multi-scale method for exposure fusion. Learning from SSIM, we choose contrast, structure, and luminance as out quality measurements. The first and the second measurements make the results more informative and the new function of the last measurement removes over-exposed and under-exposed areas. In addition, we deal the images in Lab color space, this also make our result closer to human visual system.

However, there are also some places to improve. The first one is our method only considers about luminance information, that is, it weakens color information. That is why some of areas in our result images have lower saturation. We hope that we can separately deal with L, a, b three channel to make fused images more colorful in the future. For example, just like some tone-mapping technique, fuse the luminance map then add color information to it, might be a solution.

Second, even though we find out that choosing t between 0.3 and 0.5 results in better results by our observation, there are some difference between different sequences. We will try to choose t adaptively in the future.

The last one is our performance. Exposure fusion is widely used in user applications on personal computer, digital camera and smart phone. To make users have good experience, making applications for each device be able to run in real-time is necessary. Thus, we consider generalizing our algorithm to real-time, graphic card implementation might be a good way to improve performance.

Furthermore, observe from the fused image of different sizes, we should use larger patch size for large images, or two of the quality measurement, contrast and structure, will become useless. However, larger patch size leads to longer computation time which is undesirable.

REFERENCES

[1] K. D. Ma, K. Zeng, and Z. Wang, “Perceptual Quality Assessment for Multi-Exposure Image Fusion,” *IEEE Transactions on Image Processing*, Vol. 24, No. 11, pp. 3345-3356, 2015.

[2] P. E. Debevec and J. Malik. “Recovering high dynamic range radiance maps from photographs,” *SIGGRAPH ’97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pp. 369–378, New York, NY, USA, 1997. ACM Press/Addison-Wesley Publishing Co.

[3] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, “Photographics Tone Reproduction for Digital Images,” *SIGGRAPH 2002*, 2002.

[4] R. Fattal, D. Lischinski, and M. Werman, “Gradient Domain High Dynamic Range Compression,” *SIGGRAPH 2002*, 2002.

[5] F. Durand, and J. Dorsey, “Fast Bilateral Filtering for the Display of High Dynamic Range Images”, *SIGGRAPH 2002*, 2002.

[6] P. Burt and T. Adelson. The Laplacian Pyramid as a Compact Image Code. *IEEE Transactions on Communication*, COM-31:532–540, 1983.

[7] E. Reinhard, W. Heidrich, P. Debevec, S. Pattanaik, G. Ward, and K. Myszkowski, “High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting. San Mateo”, CA, USA: Morgan Kaufmann, 2010.

[8] S. Raman and S. Chaudhuri, “Bilateral filter based compositing for variable exposure photography,” in *Proc. Eurographics*, 2009, pp. 1–4.

[9] B. Gu, W. Li, J. Wong, M. Zhu, and M. Wang, “Gradient field multi-exposure images fusion for high dynamic range image visualization,” *J. Vis. Commun. Image Represent.*, Vol. 23, no. 4, pp. 604–610, 2012.

[10] T. Mertens, J. Kautz, and F. Van Reeth, “Exposure Fusion: A Simple and Practical Alternative to High Dynamic Range Photography,” *Comput.Graph. Forum*, Vol. 28, No. 1, pp. 161–171, 2009.

[11] Z. G. Li, J. H. Zheng, and S. Rahardja, “Detail-enhanced exposure fusion,” *IEEE Transactions on Image Processing*, vol. 21, no. 11, pp. 4672–4676, 2012.

[12] S. Li, and X. Kang, “Fast multi-exposure image fusion with median filter and recursive filter,” *IEEE Trans. Consum. Electron.*, vol. 58, no. 2, pp. 626–632, May 2012.

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

[14] C. O. Ancuti, C. Ancuti, C. D. Vleeschouwer, and A. C. Bovik, "Single-Scale Fusion: An Effective Approach to Merging Images," *IEEE Transactions on Image Processing*, Vol. 26, No. 1, pp. 65-78, 2017.

[15] Q. Wang, W. Chen, and X. Lu, Z. Li, "Detail preserving multi scale exposure fusion," *Proc. IEEE Int. Conf. Image Process.*, pp. 1713-1717, Oct. 2018.