J. Vis. Commun. Image R. 31 (2015) 208–224

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Dense SIFT for ghost-free multi-exposure fusion ☆

Yu Liu [a], Zengfu Wang [a,b,*]

[a] Department of Automation, University of Science and Technology of China, Hefei 230026, China
[b] Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China

ARTICLE INFO

Article history:
Received 14 January 2015
Accepted 30 June 2015
Available online 6 July 2015

Keywords:
Multi-exposure fusion
Image fusion
Dense SIFT
Image gradient
High dynamic range imaging
Tone mapping
Quality measure
Ghosting artifacts

ABSTRACT

Due to the limited capture range of common imaging sensors, a scene with high dynamic range usually cannot be well described by a single still image because some regions in it may be under-exposed or over-exposed. In this paper, a new multi-exposure fusion method based on dense scale invariant feature transform (SIFT) is presented. In our algorithm, the dense SIFT descriptor is first employed as the activity level measurement to extract local details from source images, and then adopted to remove ghosting artifacts when the captured scene is dynamic with moving objects. Furthermore, two popular weight distribution strategies for local contrast extraction, namely, "weighted-average" and "winner-take-all" are studied in this paper. The effects of these two strategies on the fusion results are compared and discussed. Experimental results demonstrate the effectiveness of the proposed method in terms of both visual quality and objective evaluation.

1. Introduction

Common imaging sensors used in digital cameras often have very limited capture range when compared with the wide dynamic range of a natural scene. As a consequence, a single photograph only captures a certain dynamic range of the scene while tends to lose some details in both under-exposed and over-exposed regions. One way to solve this problem is utilizing high dynamic range (HDR) devices to capture as well as display real scenes. However, these devices are not popular to the general public for their high expenses. In recent years, the software-based HDR imaging technique has attracted many researchers and many effective approaches have been developed. The basic idea of these methods is that an everywhere well-exposed image can be achieved by merging a stack of low dynamic range (LDR) images which are taken by conventional LDR imaging devices like consumer cameras. Specifically, these LDR images are obtained with different exposure settings so that each of them captures a certain dynamic range of the same scene. In general, the existing HDR imaging methods can be further categorized into two main groups: tone mapping (TM)-based methods and image fusion (IF)-based methods.

In the TM-based methods, an HDR image is first reconstructed with a certain HDR reconstruction (HDR-R) technique performed over multiple LDR images of the same scene [1,2]. Although the obtained HDR images own higher fidelity than the LDR images and have been successfully used in many image and vision applications [2], most display devices in people's daily life today are not capable of displaying them. In order to fill up this gap, the TM technique is employed to generate a tone mapped image from the HDR image, which aims to simultaneously reduce the dynamic range of the HDR image and preserve its details. Thus, the TM-based methods contain two essential steps, namely, HDR reconstruction and tone mapping (HDR-R+TM). Many effective TM algorithms [3–5] have been introduced in the past few years. Reinhard et al. [3] made a compression for the luminance channel of the HDR image with a multi-scale local contrast measurement. Kuang et al. [4] presented a TM operator based on a refined image color appearance model named as iCAM06. Shan et al. [5] introduced an overlapping window-based TM approach using local linear adjustment. However, HDR reconstruction usually needs to calibrate the camera response function using some related exposure settings such as exposure time and exposure value, which are usually involved in the Exif information of the input LDR images. Unfortunately, the Exif information of a photograph is likely to be modified or lost if the picture has been artificially processed. In particular, when the LDR images are not taken by the user, the related exposure settings are often unknown. In this situation, the HDR reconstruction results may be in low quality. Since the quality of the final tone mapped result relies heavily on the reconstructed HDR image, this

☆ This paper has been recommended for acceptance by Yehoshua Zeevi.

* Corresponding author at: Department of Automation, University of Science and Technology of China, Hefei 230026, China.
E-mail addresses: liuyu1@mail.ustc.edu.cn (Y. Liu), zfwang@ustc.edu.cn (Z. Wang).

http://dx.doi.org/10.1016/j.jvcir.2015.06.021
1047-3203/

category of methods may not work well if the input exposure information is not available to users. In addition, the computational efficiency of these two-phase methods is usually not high [6].

Different from the TM-based methods which need to generate an intermediate HDR image, the IF-based methods aim to directly obtain an everywhere well-exposed image by merging the complementary information of input LDR images. Thus, the IF-based methods are usually more efficient than the TM-based methods [6]. In addition, the details of input exposure parameters are also not required, which makes the IF-based methods more convenient for ordinary users. This IF-based HDR imaging technique is also known as multi-exposure fusion, which has emerged as an active research topic in both image fusion and HDR imaging. Compared with general image fusion task [7,8], multi-exposure fusion has its own characteristics since both under-exposed and over-exposed regions in the LDR images have some distinctive properties such as extreme luminance and saturation. Although some universal fusion methods can be applied to multi-exposure fusion as well [9,10], it has significant meaning to develop more specific algorithms to pursue a better performance. Goshtasby [11] proposed a block-based multi-exposure fusion method by choosing the block which contains maximal information. Mertens et al. [12] presented an influential exposure fusion method based on weight maps. In their method, the weight map of each source image is calculated by combining three quality measures which are contrast, saturation and well-exposedness at pixel level. Moreover, to pursue a result with higher visual quality, they apply a multi-resolution-based blending approach in which the source images are decomposed into Laplacian pyramids and the weight maps are decomposed into Gaussian pyramids. Gu et al. [13] proposed a multi-exposure image fusion algorithm in the modified gradient field. Shen et al. [6] and Song et al. [14] individually introduced their probabilistic model-based method for the fusion of multi-exposure images.

The multi-exposure fusion methods [6,11–14] referred above can work well in the situation that the scene is static during the captures of multiple LDR images. However, when the scene is dynamic, which means that there exist moving objects such as walking people, running motors and windblown leaves during different captures (please note that the influence of camera motion is not included in the topic of this work), these methods may result in undesirable ghosting artifacts in the fused image. A typical example is given in Fig. 1. Fig. 1(a) shows five source LDR images with different exposures and there are some walking people in the scene during different captures. The fused result of the fusion method proposed in [12] is shown in Fig. 1(b). It can be clearly seen that serious ghosting artifacts caused by the walking people are produced in the fused image. Practically, it is often difficult to ensure the scene is absolutely static during the capture process, so developing fusion algorithms which can achieve ghost-free fusion results in dynamic scenes is of great meaning. In recent years, several fusion methods with the ability of removing ghosting artifacts have been presented [15–17]. Zhang and Cham [15] proposed a gradient-based method for multi-exposure fusion. The gradient magnitude is used to extract local contrast while the gradient direction is employed for preserving the spatial consistency to obtain a ghost-free result. An et al. [16] introduced a fusion approach based on the framework presented in [12]. They tackled the dynamic scene mainly by applying the prior of photometric relation over different source LDR images. Li and Kang [17] employed a time-domain median filter to generate a median image for ghosting artifacts removal. Besides, they used the edge-preserving recursive filter [18] to refine the final weight maps of source images. The deghosting approaches referred above all belong to the IF-based methods. Actually, more studies on deghosting in HDR imaging are carried out following the above HDR-R+TM route, in which the ghosting artifacts are removed during the HDR reconstruction process. Please refer to [19] for a comprehensive survey on ghost removal methods in HDR imaging. Recently, some new deghosting methods [20,21] have been proposed in this literature, and there are also some commercial software packages available to obtain ghost-free HDR images, such as Photoshop and Photomatix. Hadziabdic et al. [22] provided a comparison among these latest deghosting approaches through subjective psychophysical experiments. At the present time, the deghosting methods based on HDR-R+TM way are much more studied, leading to state-of-the-art performance [19,22]. However, this category of methods shares the common shortcomings of the TM-based HDR imaging methods. These methods usually rely heavily on an accurate estimation of camera response function and require a TM operator to make the HDR image displayable. Thus, we believe it is of great significance to put more efforts into the study of IF-based deghosting methods.

In multi-exposure fusion literature, the fusion method [12] takes up an important position. As mentioned above, three quality measures, namely, contrast, saturation and well-exposedness are used to construct the pixel-level weight maps for source LDR images. Specifically, a weight term is calculated for each quality measure, and then a weight map is obtained by multiplying all of its corresponding terms. This weight term-based fusion scheme has a great influence on many later fusion methods [15–17,23,24], in which different weight terms, calculation approaches, and post-processing techniques for weight map refinement have been proposed. This fusion scheme can usually achieve good performance and owns advantages in terms of simplicity, efficiency and flexibility. In particular, a weight term used for removing ghosting artifacts in dynamic scene can also be designed and combined into the fusion framework [15–17,23,24]. Among all these methods, the local contrast information constructs an important weight term. Since a well-exposed local region usually contains more spatial details than the corresponding under- or over-exposed region, it is necessary to design a local contrast measure to extract the spatial details in the scene from source LDR images. In previous publications, Laplacian filtering [12,16,17,23,24] and gradient magnitude [15] have been employed to measure the contrast of each pixel in source images. With a certain local contrast measure, a weight distribution strategy is designed to assign the weights for each source image, i.e., to calculate the local contrast weight term. In general, there are two popular weight distribution strategies: "weighted-average" and "winner-take-all". The first one implies that each source image gets a weight value according to its response of contrast measure, while the second means only the source image owning the strongest response (highest measurement) gains the total weights. The selection of distribution strategy has a great impact on the final fused result. However, in previous multi-exposure fusion research, to the best of our knowledge, only one of the above two strategies is employed in a multi-exposure fusion method (the "weighted-average" strategy is applied in [12,15,16,23,24], and the "winner-take-all" strategy is applied in [17]), and there is no study that focuses on the comparison between them.

In this paper, we propose a ghost-free multi-exposure fusion method based on dense scale invariant feature transform (SIFT) [25] with the weight term-based scheme. In our algorithm, the dense SIFT descriptor is used for both local contrast extraction and ghosting artifact removal when the images are taken in dynamic scenes. Furthermore, the two distribution strategies mentioned above are tested and compared. Fig. 1(c) and (d) shows the fused results of the proposed method with the "weighted-average" and "winner-take-all" strategies, respectively. On one hand, we can see that both of the two results exhibit the spatial details in the scene well and the ghosting artifacts produced in Fig. 1(b) are

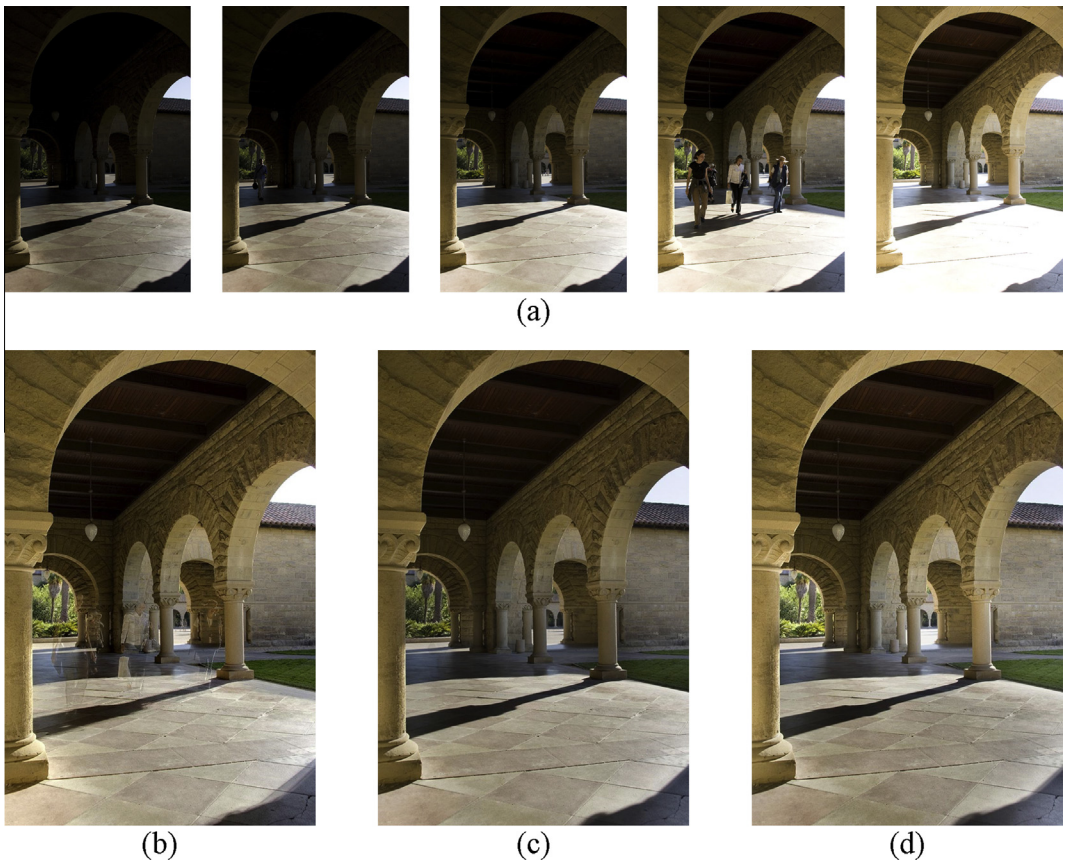

Fig. 1. A dynamic multi-exposure fusion example. (a) Source LDR images. (b) Result by Mertens et al. [12]. (c) Our result with "weighted-average" strategy. (d) Our result with "winner-take-all" strategy. Data courtesy of Orazio Gallo.

successfully eliminated. On the other hand, there exists clear difference between the two results. It can be seen that the result obtained with "winner-take-all" strategy has higher contrast in some regions, but it contains some color distortions in some regions such as the shadow at lower right corner. The main contributions of this paper are summarized as the following three points.

1. As a local feature of images, dense SIFT is first introduced into multi-exposure fusion. Two quality measures based on dense SIFT are designed for local contrast extraction and ghosting artifact removal, respectively.
2. A new multi-exposure fusion method based on dense SIFT is proposed. The method can be used to fuse LDR images taken in both static and dynamic scenes.
3. Two popular weight distribution strategies ("weighted-aver age" and "winner-take-all") for local contrast extraction in multi-exposure fusion are studied. Their effects on the fusion results are compared and discussed.

The rest of this paper is organized as follows. In Section 2, we present the basic motivation of dense SIFT for multi-exposure fusion. Section 3 describes the proposed fusion algorithm in detail. The experimental results and discussions are presented in Section 4. Finally, Section 5 concludes the paper.

2. Dense SIFT for multi-exposure fusion

Dense SIFT was introduced into the field of multi-focus image fusion in our recent work [26]. This paper aims to further exhibit the potential of dense SIFT for multi-exposure fusion in both static and dynamic scenes. In this section, we first give a brief introduction to dense SIFT. Then, the basic idea of dense SIFT-based multi-exposure fusion is presented. Finally, the advantages of dense SIFT over image gradient [15] for multi-exposure fusion are discussed.

The well-known SIFT descriptor proposed by Lowe [27] achieves great success in various computer vision applications. For a detected interest point, its SIFT descriptor is generated by characterizing its local gradient information. Thus, the SIFT descriptor contains underlying salient information, which can be used to design activity level measurement [28] for image fusion. However, the locations of interest points are sparse and uncertain. In the field of pixel-level image fusion, the activity level measurement must be assigned to each pixel or at least each local block, so the SIFT descriptor [27] cannot be directly employed. Recently, Liu et al. [25] presented the dense SIFT descriptor for image registration and face recognition. In dense SIFT, a feature descriptor can be extracted for each pixel in an image and the process for detecting interest points in [27] is not required. The calculation of a descriptor for a pixel is similar to the approach proposed in SIFT algorithm [27]. Specifically, the local region around the pixel is first divided into several cells. Then, an orientation histogram with several bins is employed to characterize the gradient information in each cell. For each cell, its histogram is obtained by accumulating the gradient magnitude of each pixel in it into a corresponding bin which is selected according to the pixel's gradient orientation. Accordingly, a feature descriptor is formed for each pixel. At last, the obtained descriptors usually need to be normalized with the approach in [27] for some future applications such as feature matching. Unlike the SIFT descriptor, the dense SIFT descriptor is neither scale nor rotation invariant, but fortunately these two invariant properties are not necessary for image fusion since the

source images are assumed to be captured via a tripod or pre-registered using image registration techniques. Actually, SIFT descriptor has been employed for multi-exposure image registration by Tomaszewska and Mantiuk [29], while this paper first introduces dense SIFT into the field of multi-exposure fusion.

The main advantage of dense SIFT used for image fusion is that it simultaneously addresses two key issues of image fusion: the activity level measurement of each source image and the local similarity between multiple source images. The former one, which aims to extract the local contrast information in multiple source images, plays a pivotal role in a variety of image fusion scenarios such as multi-focus, multi-modal and multi-exposure. The latter one is of great importance when the source images are not well registered or there exist moving objects in the scene. In this paper, the basic idea of dense SIFT-based multi-exposure fusion can be summarized as the following two points. First, the unnormalized dense SIFT descriptor is used to extract the local contrast information of source images. This point is similar to that in [26] for multi-focus image fusion, i.e., the sum of all the elements in an unnormalized descriptor vector, which quantifies the extent of intensity variation within a local image patch, is used as the activity level measurement to extract local details. For the local patches in different source images with the same position, a higher value of this measure indicates more details are contained. However, the fusion scheme is much simpler here than [26]. For multi-focus image fusion, local contrast is the most important issue since it directly reflects the focus/clarity level, so we designed a complex sliding window-based approach to fuse multi-focus images in [26]. For multi-exposure fusion, differently, more issues as mentioned above should be considered and local contrast is just one of them. Thus, it is not necessary to adopt too complicated approach at the sacrifice of computational efficiency. Second, the normalized dense SIFT descriptor is used to obtain ghost-free results when the source images are captured in dynamic scenes. In the multi-focus image fusion work [26], the normalized SIFT descriptor is utilized to improve the fusion quality over object edges by adjusting the mis-registered pixels between different source images. For each pixel in one source image, a corresponding neighbor region in other source images are used for matching. The situation in multi-exposure fusion is largely different. The target here is to remove the ghosting artifacts caused by moving objects which may cover a large area in the image, so the neighbor matching approach used in [26] loses its effectiveness. In this work, the descriptors of the pixels with the same location in multiple source images are used to measure the local similarity of image contents for ghost removal. This is feasible because the normalized SIFT descriptor is generally invariant to illuminance change so long as the corresponding region is not under-exposed or over-exposed. When a moving object occurs, no matter in which exposure, the descriptor usually varies significantly with respect to the background. This is the basic motivation that we employ dense SIFT for ghost-free multi-exposure fusion, while further discussions will be given after the experimental results are exhibited.

It is worthwhile to notice that the dense SIFT-based quality measures share some similarity with the image gradient-based measures proposed in [15] since the dense SIFT descriptor is essentially obtained based on image gradient. However, compared with directly using gradient magnitude and direction, the dense SIFT-based approach owns the following two advantages.

1. *Local contrast extraction.* The dense SIFT descriptor of a pixel is obtained by utilizing the gradient information of all pixels in its local patch, so the related activity level measurement is more reliable and robust to noise than using gradient magnitude.
2. *Ghosting artifact removal.* The dense SIFT descriptor characterizes the gradient information of a pixel on several orientations. Furthermore, with the normalization technique, the descriptor is more robust to illumination variations than directly using the gradient direction. Thus, the dense SIFT descriptor of a pixel is much more detailed and accurate than its gradient direction when describing its corresponding local patch.

3. Detailed fusion method

3.1. Overview

In this section, the dense SIFT-based multi-exposure fusion algorithm is presented in detail. Fig. 2 shows the schematic diagram of the proposed method. In this work, the input source images are assumed to be captured via a tripod or pre-registered with some image alignment techniques such as [29,30]. In our method, for each source image, we first construct three weight terms which are local contrast, exposure quality and spatial consistency (when the images are captured in static scenes, the spatial consistency term is not required). Then, a weight map is estimated for each source image by combining its weight terms. After that, the estimated weight maps of all source images are refined and normalized. Finally, the resulting fused image is obtained by calculating the weighted sum of source images at each pixel.

3.2. Weight term construction

3.2.1. Local contrast

Let $I_i, i = 1, 2, \ldots, N$ denote the source images. In our algorithm, color source images are firstly converted to gray ones by $\hat{I}_i = 0.299I_i^r + 0.587I_i^g + 0.114I_i^b$ [31], where I_i^r, I_i^g and I_i^b are the red, green and blue channel of I_i, respectively (when I_i is a gray image, just keep $\hat{I}_i = I_i$). The unnormalized dense SIFT map $\mathbf{S}_i(x,y)$ of $\hat{I}_i(x,y)$ is obtained by

$$\mathbf{S}_i(x,y) = \mathrm{DSIFT}\left(\hat{I}_i(x,y)\right), \tag{1}$$

where $\mathrm{DSIFT}(\cdot)$ denotes the operator which aims to calculate the unnormalized dense SIFT map for an input image. Please refer to [25,27] for more details about the calculation of dense SIFT. For a certain pixel located at (x,y), $\mathbf{S}_i(x,y)$ is a descriptor vector. To reduce memory consumption, a descriptor is generated by applying a 2×2 cell array and an 8-bin orientation histogram in each cell, so the dimension of each descriptor vector is 32.

As mentioned in Section 2, the sum of all the elements in an unnormalized descriptor is used to measure the activity level of the corresponding pixel. Since all the elements in a SIFT descriptor are not negative, the activity level map of $\hat{I}_i(x,y)$ can be represented as the l_1 norm of $\mathbf{S}_i(x,y)$ at each pixel:

$$A_i(x,y) = \|\mathbf{S}_i(x,y)\|_1. \tag{2}$$

Then, two weight distribution strategies mentioned above, which are "weighted-average" and "winner-take-all", are employed to construct the local contrast weight term $\hat{A}_i(x,y)$.

(i) "weighted-average" strategy:

$$\hat{A}_i(x,y) = A_i(x,y), \tag{3}$$

(ii) "winner-take-all" strategy:

$$\hat{A}_i(x,y) = \begin{cases} 1, & A_i(x,y) = \max\{A_i(x,y), i = 1, 2, \ldots, N\}, \\ 0, & \text{otherwise}. \end{cases} \tag{4}$$

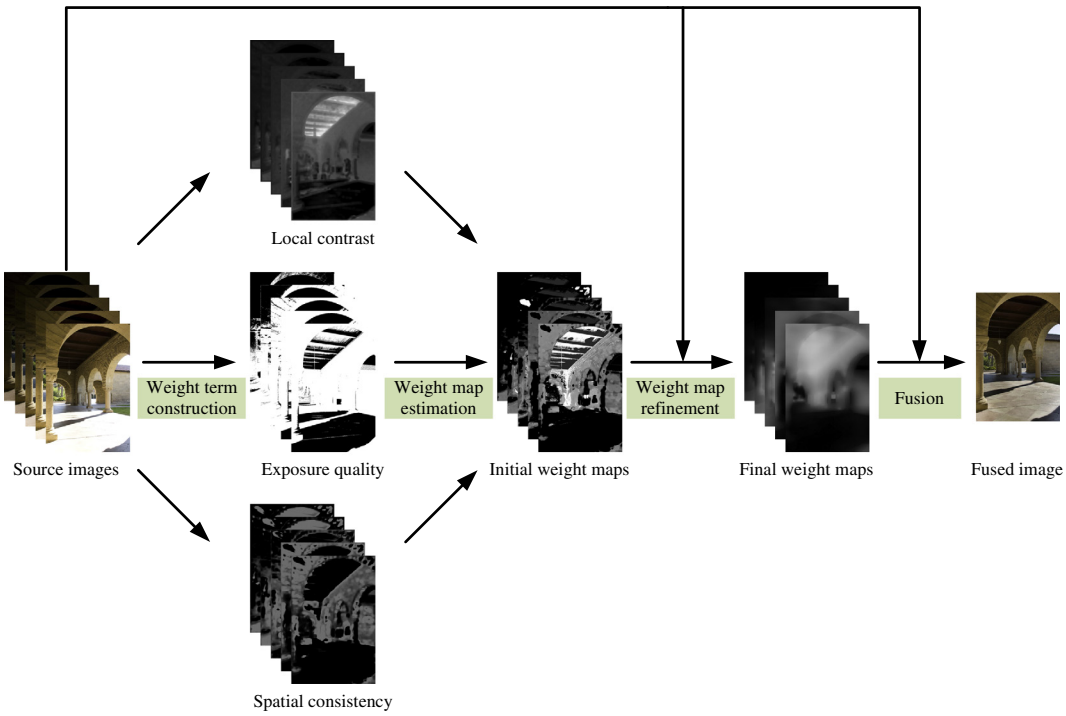

Fig. 2. Schematic diagram of the proposed multi-exposure fusion method.

3.2.2. Exposure quality

We use the brightness of a pixel to roughly measure its exposure quality. The basic assumption is that under-exposed and over-exposed pixels usually have very weak and strong brightness, respectively. The well-exposed pixels in each source image are detected to construct its exposure quality weight term:

$$B_i(x,y) = \begin{cases} 1, & \alpha < \hat{I}_i(x,y) < 1-\alpha, \\ 0, & \text{otherwise}. \end{cases} \tag{5}$$

$B_i(x,y) = 1$ indicates the related pixel is well-exposed. The parameter $\alpha \in [0,1]$ determines the well-exposed range (the image range is normalized to $[0,1]$). This quality measure can effectively remove the influence of under-exposed or over-exposed regions on the fused image and has been used in many multi-exposure fusion methods such as [15,17]. The α is also set to 0.1 in our fusion method.

3.2.3. Spatial consistency

When the source images are captured in a dynamic scene which contains moving objects like walking people, the fused image may suffer from ghosting artifacts if only the above two terms are considered. To solve this problem, we present a spatial consistency weight term based on normalized dense SIFT descriptor. With $\mathbf{S}_i(x,y)$, the corresponding normalized dense SIFT map $\mathbf{S}_i^n(x,y)$ is obtained by

$$\mathbf{S}_i^n(x,y) = \text{Normalization}(\mathbf{S}_i(x,y)), \tag{6}$$

where Normalization$(\cdot)$ denotes the SIFT descriptor normalization operator [27].

For each pixel location (x,y), the Euclidean distance between $\mathbf{S}_i(x,y)$ and $\mathbf{S}_j(x,y)$ is used to measure the local similarity between $I_i(x,y)$ and $I_j(x,y)$. For the sake of computational efficiency, we apply the square of Euclidean distance:

$$d_{ij}(x,y)^2 = \left\| \mathbf{S}_i^n(x,y) - \mathbf{S}_j^n(x,y) \right\|_2^2. \tag{7}$$

In order to increase its robustness to noise, the window-based scheme is adopted. The local similarity map between $I_i(x,y)$ and $I_j(x,y)$ is calculated by

$$D_{ij}(x,y) = \frac{\sum_{p=-r}^{r}\sum_{q=-r}^{r} d_{ij}(x+p,y+q)^2}{(2r+1)^2}, \tag{8}$$

where r is the radius of the window and is fixed at nine. At last, the spatial consistency weight term for each source image is constructed by

$$C_i(x,y) = \sum_{j=1,j\neq i}^{N} \exp\left(-\frac{D_{ij}(x,y)}{2\sigma_d^2}\right), \tag{9}$$

where the standard deviation σ_d controls the impact of $D_{ij}(x,y)$ on $C_i(x,y)$, and is normally set to 0.05 in the method. When the pixel (x,y) in image I_i belongs to a moving object, the values of $D_{ij}(x,y)$ for all j tend to increase. As a result, the value of $C_i(x,y)$ will decrease, leading to a smaller weight of image I_i at pixel (x,y).

3.3. Weight map estimation

The goal of this step is to obtain an initial weight map for each source image by combining the information contained in its three weight terms, namely, local contrast $\hat{A}_i(x,y)$, exposure quality $B_i(x,y)$ and spatial consistency $C_i(x,y)$. The straightforward approach is to multiply these three terms one by one [12]. However, the pixels of the same location in all $B_i(x,y), i = 1,2,\ldots,N$ may be zero at the same time. In this situation, the information in either $\hat{A}_i(x,y)$ or $C_i(x,y)$ cannot be injected into the weight maps. To solve this problem, we first multiply $B_i(x,y)$ by $C_i(x,y)$ for each source image and then normalize the obtained results for all source images as follows:

$$T_i(x,y) = \begin{cases} B_i(x,y), & \text{static scene}, \\ B_i(x,y) \times C_i(x,y), & \text{dynamic scene}, \end{cases} \tag{10}$$

$$\hat{T}_i(x,y) = \frac{T_i(x,y) + \varepsilon}{\sum_{i=1}^{N}(T_i(x,y) + \varepsilon)}, \tag{11}$$

where ε is a small positive value (e.g., 10^{-25}) and the spatial consistency term is not involved when the scene is static. After normalization with Eq. (11), for a pixel located at (x,y), all $\hat{T}_i(x,y)$ will be $1/N$ if $B_i(x,y) = 0, i = 1,2,\ldots,N$, so the local contrast information in $\hat{A}_i(x,y)$ can be extracted. Finally, the initial weight map $W_i(x,y)$ of each source image is calculated by

$$W_i(x,y) = \hat{A}_i(x,y) \times \hat{T}_i(x,y). \tag{12}$$

3.4. Weight map refinement

As shown in Fig. 2, the initial weight maps obtained above are usually noisy and discontinuous, so they need to be refined before used for the final fusion. Some recent proposed edge-preserving techniques such as the joint bilateral filter [32], guided filter [33] and recursive filter [18] can be used to accomplish the task, in which the source image is served as the joint/guided image to ensure that pixels from the same objects have similar weights. Among them, the recursive filter owns real-time efficiency and has achieved success in the multi-exposure fusion method [17]. Thus, we choose recursive filter to refine the initial weight maps in our method. The refined weight map $W_i^r(x,y)$ of $W_i(x,y)$ can be represented as

$$W_i^r(x,y) = \mathrm{RF}(W_i(x,y), I_i(x,y)), \tag{13}$$

where $\mathrm{RF}(\cdot,\cdot)$ denotes the recursive filter operation, $W_i(x,y)$ is the input image to be filtered and $I_i(x,y)$ is the joint image.

Then, the refined weight maps are normalized to guarantee that the sum of all the weight maps is one at each pixel location. Thus, the final weight map of each source image is obtained by

$$\hat{W}_i(x,y) = \frac{W_i^r(x,y) + \varepsilon}{\sum_{i=1}^{N}(W_i^r(x,y) + \varepsilon)}. \tag{14}$$

3.5. Fusion

At last, with the obtained final weight maps, the resulting fused image I_F is calculated as the weighted sum of source images at each pixel:

$$I_F(x,y) = \sum_{i=1}^{N} \hat{W}_i(x,y) \times I_i(x,y). \tag{15}$$

4. Experiments

4.1. Experimental setup

In our experiments, six multi-exposure image sequences are used to test the effectiveness of the proposed fusion method. Fig. 3 shows two images under different exposures in each sequence. More details about each sequence such as the name, the spatial resolution and the number of source images are described in the caption of Fig. 3. Among the six test sequences, three (Fig. 3(a)–(c)) are taken in static scenes and the other three (Fig. 3(d)–(f)) are taken in dynamic scenes. Furthermore, the first five sequences (Fig. 3(a)–(e)) are standard test sets, and we capture the last one (Fig. 3(f)) with a Nikon D3200 DSLR camera via a tripod. Please notice that the test image sequences cover indoor scenes, outdoor scenes, vegetation scenes, architectural scenes, etc.

The proposed DSIFT-based fusion method has a free parameter: the scale factor, namely, the size of each pixel's neighborhood for DSIFT calculation [25]. In all the following experiments unless otherwise specified, the scale factor is fixed at 16. The parameters with respect to recursive filter are set as the reported optimal values. For simplicity, we denote the algorithm which applies the "weighted-average" strategy in the construction of local contrast weight term as DSIFT-1, and the algorithm with the "winner-take-all" strategy is denoted as DSIFT-2. *The MATLAB implementation of the proposed multi-exposure fusion method will be made available on http://home.ustc.edu.cn/~liuyu1.*

Four existing multi-exposure fusion methods, i.e., the exposure fusion (EF) method [12], the generalized random walks (GRW) method [6], the fusion method named as FMMR [17], and the image gradient (IG)-based fusion method [15] are employed to make comparisons. The EF method and the GRW method can only tackle image sequences captured in static scenes. The code of the EF method is available on website [34]. The executable file of the GRW method is provided by its first author Shen. The FMMR method and the IG method can obtain state-of-the-art results in multi-exposure fusion for dynamic scenes. The code of the FMMR method is downloaded from Kang's homepage [35]. The IG method is implemented strictly based on [15]. It should be noted that the proposed DSIFT method in this paper shares the same fusion scheme with the IG method [15]. The difference between them is the construction of local contrast term and spatial consistency term. The main purpose is to fairly verify the advantage of DSIFT-based quality measures over the IG-based measures used in [15]. In fact, the proposed DSIFT-based measures can be universally employed in weight term-based multi-exposure fusion

Fig. 3. Source multi-exposure image sequences used in the experiments. (a) "Memorial", 464 × 696 × 16 (the meaningless blue boundaries are cut away from the original source images), data courtesy of Paul Debevec. (b) "B-House", 1025 × 769 × 9, data courtesy of Dani Lischinski. (c) "Garage", 348 × 222 × 6, data courtesy of Shree K. Nayar. (d) "Arch", 669 × 1024 × 5, data courtesy of Orazio Gallo. (e) "Forest", 1024 × 683 × 4, data courtesy of Orazio Gallo. (f) "Campus", 1200 × 800 × 4.

scheme. One may notice that the IG method [15] adopted the joint bilateral filter [32] for weight map refinement. To make a fair comparison, we replace it with the recursive filter, which has been verified to be capable of obtaining state-of-the-art edge-preserving performance [18]. Actually, we experimentally find that this replacement can slightly improve the performance of the original IG method [15]. Furthermore, the recursive filter is much more efficient than the joint bilateral filter. The two weight distribution strategies are both tested in the IG method as well. As the notation used in the DSIFT method, the IG method using the "weighted-average" strategy is denoted as IG-1 and the method using the "winner-take-all" strategy is denoted as IG-2. For all the above fusion methods, the default parameter settings reported in related publications are used in all experiments of this paper.

In addition, the proposed method is also compared with three TM-based methods which are the iCAM06 method [4], the linear windowed (LW) method [5] and the default TM method integrated in Photomatix, which is a commercially available software package for HDR imaging [36]. One important advantage of Photomatix over traditional HDR imaging methods like [1] is that it can effectively remove ghost (an optional function for users) when generating the HDR image. The results reported in [22,37] demonstrate that Photomatix is competitive with several latest HDR deghosting methods and can generally produce state-of-the-art results. Furthermore, the well-designed interface makes a very convenient utilization of Photomatix. In our experiments, the intermediate HDR images for the iCAM06 method and the LW method are generated by Photomatix Pro (version 4.2.5), and the default TM results of Photomatix are also used for comparison. The implementations of the iCAM06 method and the LW method are available on websites [38,39], respectively. The default parameter settings in these two methods are used in all experiments.

4.2. Experimental results and discussions

This subsection exhibits the experimental results and presents some relevant discussions. To make a better comparison, a close-up is given for each sequence.

The experimental results of different methods on the classical "Memorial" sequences are shown in Fig. 4. This image sequence was captured inside a memorial. It is very difficult to simultaneously capture the details of both the inside walls and the window decorations. It can be seen that the LW method suffers from severe color distortion, and the other two TM-based methods tend to have low brightness. It does not indicate that these TM-based methods cannot obtain good performance because the TM-based methods rely heavily on the HDR image obtained in advance. However, this disadvantage indeed limits the usefulness of these methods. The results of the EF and GRW methods are in high quality over the indoor walls, but suffer from serious over-exposure in the window regions (see the close-ups in Fig. 4(d) and (e)). The results of the FMMR, IG-2 and DSIFT-2 methods have higher contrast than the results of the IG-1 and DSIFT-1 methods in some regions (see the ceiling regions in Fig. 4(f)–(j)), but the results obtained with

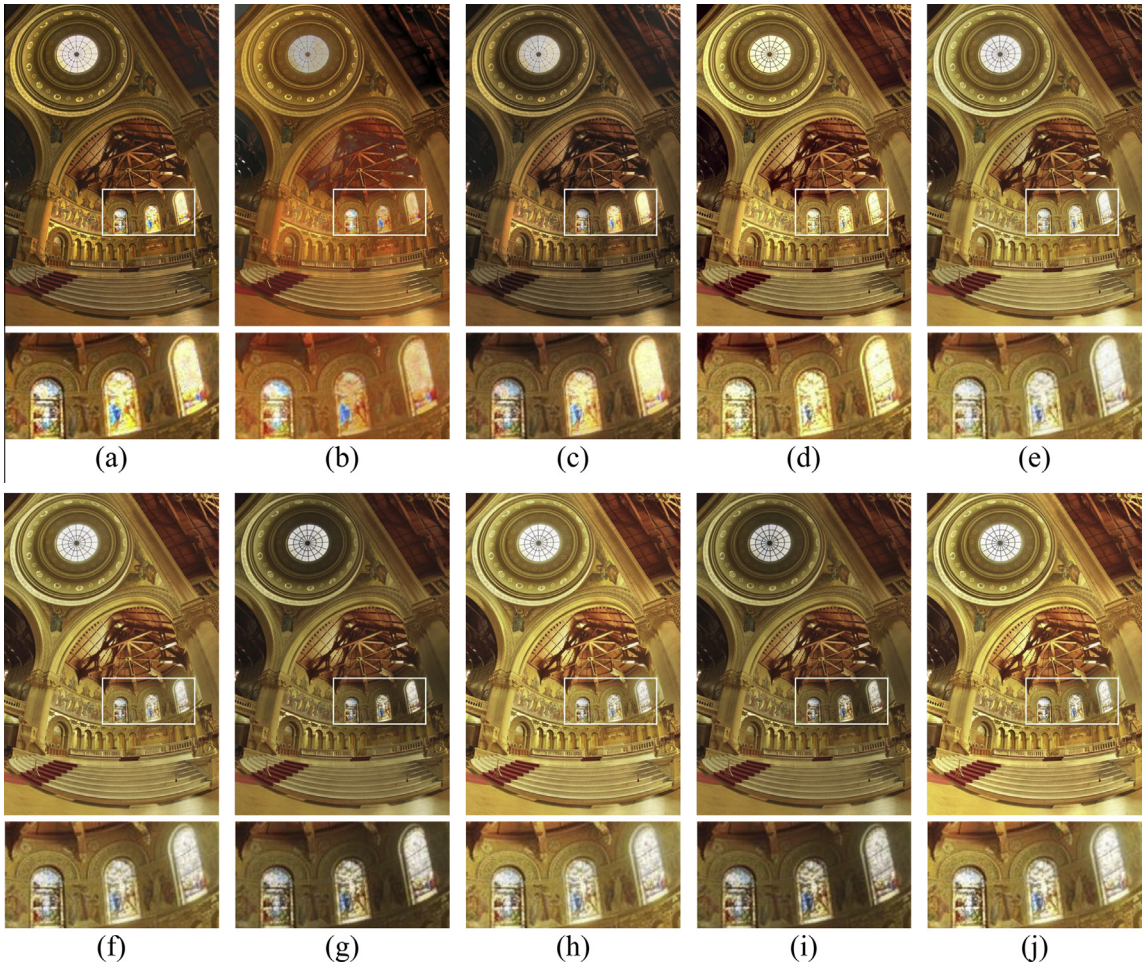

Fig. 4. Performance comparison of different methods on the "Memorial" image sequence. (a) iCAM06. (b) LW. (c) Photomatix. (d) EF. (e) GRW. (f) FMMR. (g) IG-1. (h) IG-2. (i) DSIFT-1. (j) DSIFT-2.

"winner-take-all" strategy suffer from over-exposure in some regions (see the ceiling above the cropped region). We believe it is a matter of opinion which result owns the highest visual quality in this example.

The results on the "B-House" sequences are shown in Fig. 5. This sequence was captured inside a room, but with an outdoor garden in the scene. For the TM-based methods, the situation is similar to that in Fig. 4. The EF and GRW methods both suffer form over-exposure in the outdoor regions (especially for the EF method), and some indoor regions are under-exposed (see the close-ups in Fig. 5(d) and (e)). Dark halo occurs in the wall regions in the results of the IG-1 and DSIFT-1 methods, which degrades the visual quality to a large extent. The situation is much improved by the "winner-take-all" strategy used in the FMMR, IG-2 and DSIFT-2 methods, in particular for the DSIFT-2 method (see the wall regions in Fig. 5(f), (h) and (j)). Furthermore, the DSIFT-2 method extracts more local details than other fusion methods (see the close-ups and ceiling regions in Fig. 5(d)–(j)).

Fig. 6 shows the results on the "Garage" image sequence, which was taken in an outdoor scene with a garage existing in the dark. It

Fig. 5. Performance comparison of different methods on the "B-House" image sequence. (a) iCAM06. (b) LW. (c) Photomatix. (d) EF. (e) GRW. (f) FMMR. (g) IG-1. (h) IG-2. (i) DSIFT-1. (j) DSIFT-2.

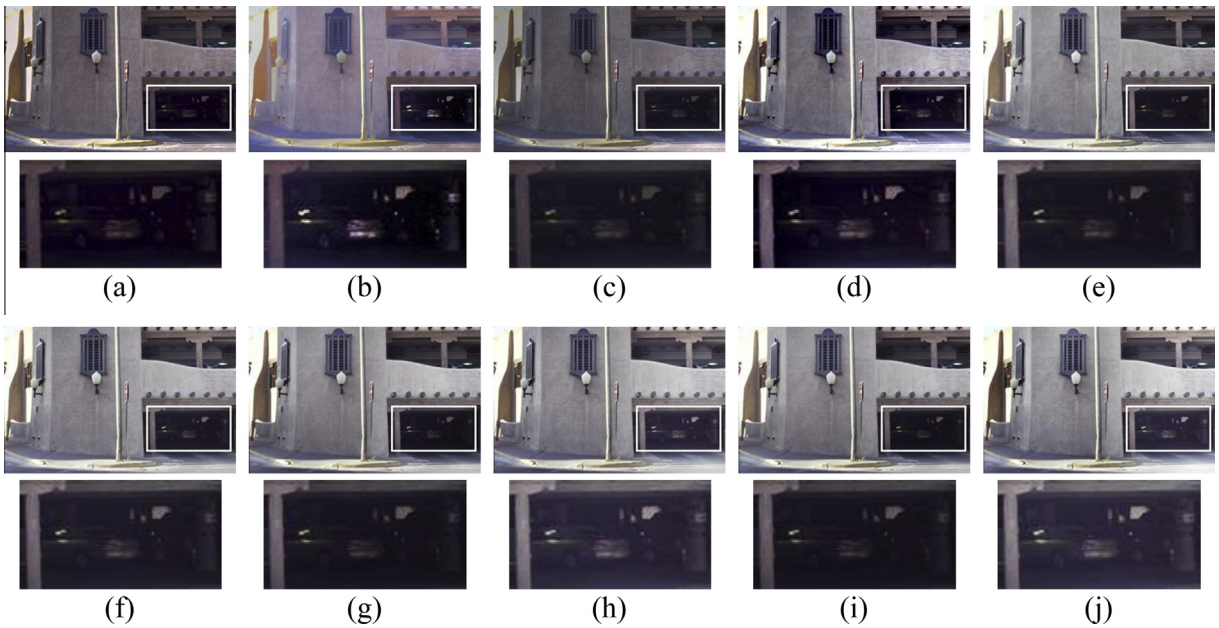

Fig. 6. Performance comparison of different methods on the "Garage" image sequence. (a) iCAM06. (b) LW. (c) Photomatix. (d) EF. (e) GRW. (f) FMMR. (g) IG-1. (h) IG-2. (i) DSIFT-1. (j) DSIFT-2.

can be seen that color distortion still exists in the results of TM-based methods. The result of the EF method loses some important spatial details (see the window in Fig. 6(d)). In addition, the information in the garage region is not well merged (see the close-up in Fig. 6(d)). The GRW, IG-1, DSIFT-1 methods have good performances in the outdoor regions, but these results suffer from under-exposure in the garage region as well (see the close-ups in Fig. 6(e), (g) and (i)). An improvement of the fusion quality in the garage region is made by the FMMR method, but the effect is still not satisfactory when considering the source inputs. The results of the IG-2 and DSIFT-2 methods provide the highest visual quality in the garage region (see the close-ups in Fig. 6(d)–(j)). Furthermore, the FMMR, IG-2 and DSIFT-2 methods extract more details than the IG-1 and DSIFT-1 methods (see the external wall of the building in Fig. 6(f)–(j)).

The experimental results on the "Arch" image sequence are shown in Fig. 7. Color distortion is very obvious in the result of the LW method. The results provided by the iCAM06 and Photomatix methods have high visual quality. An interesting observation is that the ceiling regions in Fig. 7(a) and (c) both have higher contrast than the same regions in all of the five source images (see Fig. 1(a)). Obviously, it is impossible for a weighted sum based exposure fusion method to achieve this goal. The EF and GRW methods are not competent for the fusion task in dynamic scenes. As shown in Fig. 7(d) and (e), the ghosting artifacts are very serious in the related fusion results. The results of FMMR, IG-2 and DSIFT-2 methods have relatively high contrast in some regions among IF-based methods, but the shadows are over-exposed (see the close-ups in Fig. 7(f), (h) and (j)). Moreover, the edges of the shadow suffer from ghosting artifacts in these three results (please notice that there exists movement of the shadow region over different source images). The IG-1 and DSIFT-1 methods handle the shadow regions much better, but the obtained results have relatively low contrast in some regions (see the ceiling of the building in Fig. 7(g) and (i)). Compared with the IG-1 method, the DSIFT-1 method obtain better performance in terms of ghosting artifact removal. To make a clearer comparison, Fig. 8 provides another two close-ups from the results of these two methods. We can see that the DSIFT-1 method outperforms the IG-1 method over several regions, such as the middle lady's face, the circular columns of the architecture, and the right person's contour.

Fig. 9 shows the performances of different methods on the "Forest" sequence. The characteristics of the first five results shown in Fig. 9 are very similar to those in Fig. 7. We mainly focus on the results provided by the last five fusion methods. The FMMR, IG-2 and DSIFT-2 methods also do not perform well in merging the shadow regions, and situation here is more obvious since this vegetation scene is full of trees and grasses. As a result, the grassy regions shown in Fig. 9(f), (h) and (j) seem to be very unnatural, which degrades the visual quality to a great extent. The performances of the IG-1 and DSIFT-1 methods are much improved in fusing shadow regions. These two results obtained with the "weighted-average" strategy are more natural for human visual perception. Furthermore, for either IG or DSIFT based methods, the one using the "weighted-average" strategy has a better performance on ghost removal than the one using the "winner-take-all" strategy (see the close-ups in Fig. 9(g)–(j)). Particularly, the DSIFT-1 method outperforms the IG-1 method in terms of removing ghosts (see the close-ups in Fig. 9(g) and (i)).

Fig. 10 shows the results on our "Campus" image sequence, in which four source images with different exposure values are contained. In the captured scene, in addition to the pedestrian, there exist slight motions of the branches and leaves between arbitrary two source images caused by the wind. As shown in Fig. 10(a)–(c), the results of three TM-based methods are all suffer from severe color distortion. The EF and GRW methods cannot obtain ghost-free results (see the close-ups in Fig. 10(d) and (e)). In this example, it is worthwhile to notice that the IG-1 and IG-2 methods also fail in removing ghosting artifacts (see the close-ups in Fig. 10(g) and (h)). By contrast, the ghost removal performances of the DSIFT-1 and DSIFT-2 methods are significantly improved. The FMMR method also does well in removing ghosts, but it loses some important details (see the light spots on the ground in Fig. 10(f)). It should be noted that none of these fusion methods obtains a satisfactory result in terms of the windblown branches and leaves. The fusion results of the IG-1 and DSIFT-1 methods suffer from serious blurring artifacts. The situation of blur is relatively

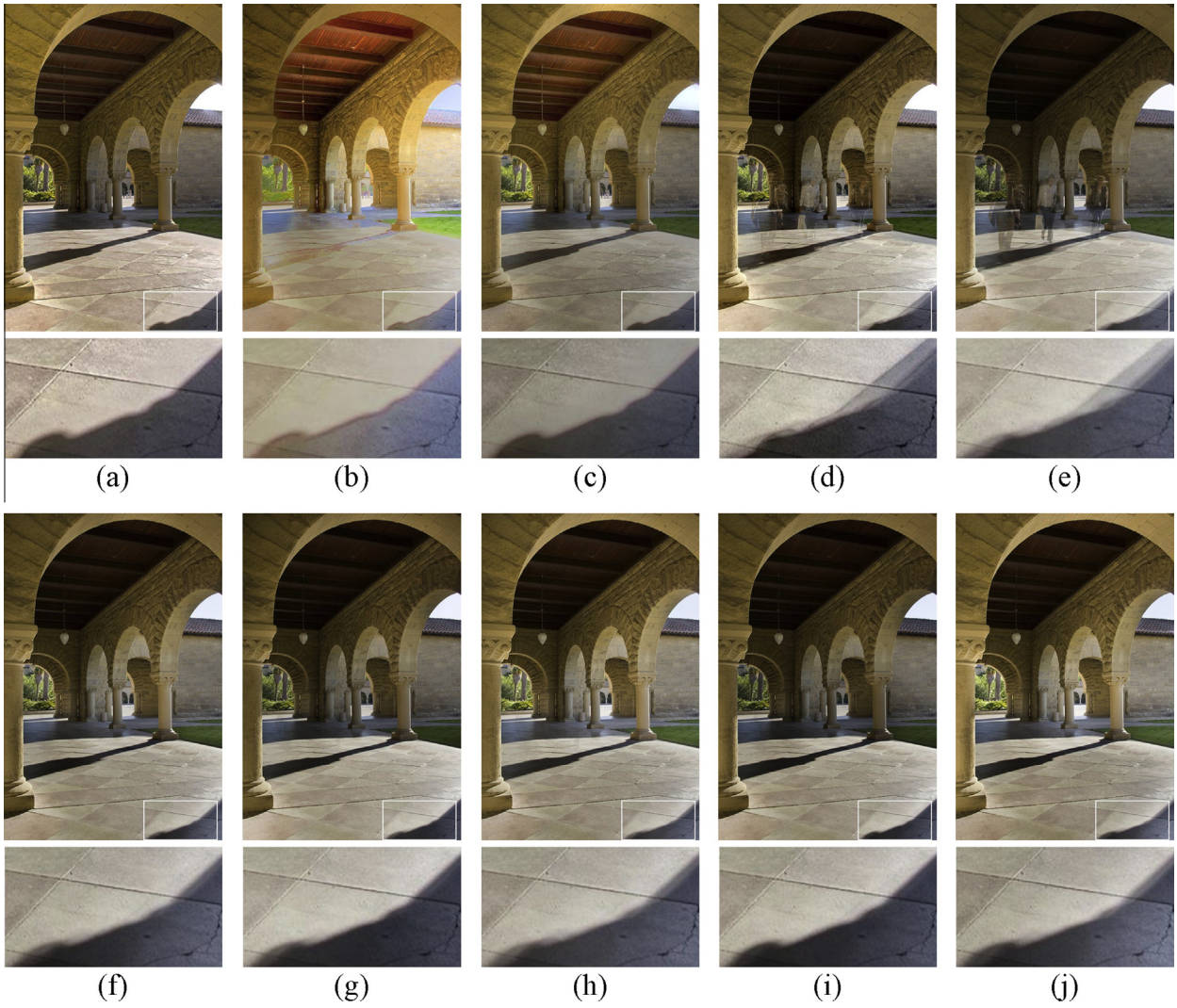

Fig. 7. Performance comparison of different methods on the "Arch" image sequence. (a) iCAM06. (b) LW. (c) Photomatix. (d) EF. (e) GRW. (f) FMMR. (g) IG-1. (h) IG-2. (i) DSIFT-1. (j) DSIFT-2.

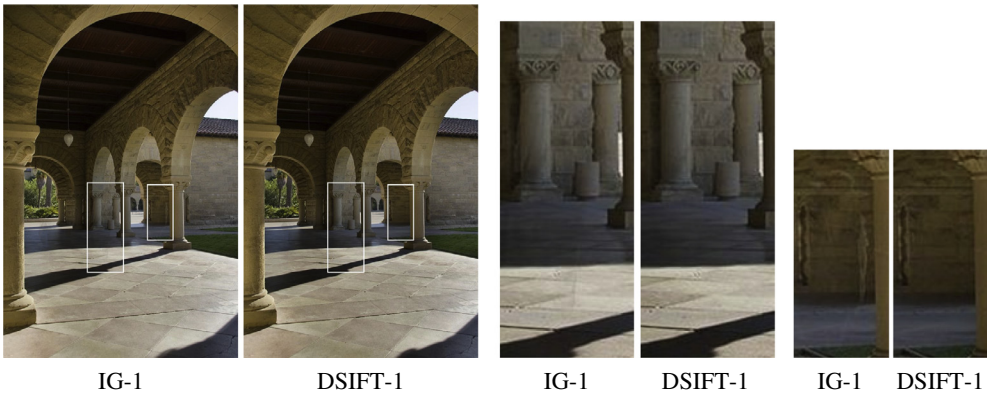

Fig. 8. Performance comparison of the IG-1 and DSIFT-1 methods on ghost removal.

better in the fused images provided by the FMMR, IG-2 and DSIFT-2 methods. However, the color of these regions seems to be too bright, which may make the visual quality even worse than the results of the IG-1 and DSIFT-1 methods.

At last, we summarize the above results and make some discussions. First, with either the "weighted-average" or "winner-take-all" strategy, the results on six image sequences indicates that the proposed DSIFT-based fusion method generally outperforms the IG-based method on both local contrast extraction and ghosting artifact removal, especially for the latter one. Second, the characteristics of two weight distribution strategies for local contrast can be summarized as follows.

Fig. 9. Performance comparison of different methods on the "Forest" image sequence. (a) iCAM06. (b) LW. (c) Photomatix. (d) EF. (e) GRW. (f) FMMR. (g) IG-1. (h) IG-2. (i) DSIFT-1. (j) DSIFT-2.

1. The "weighted-average" strategy is often a better choice when the scene is abundant in textures, especially when containing many trees and grasslands (see the "Forest" and "Campus" sequences). Furthermore, when the scene is dynamic, a fusion method using this strategy usually has a slight advantage over the same method using the "winner-take-all" strategy in terms of removing ghosting artifacts (see the "Arch" and "Forest" sequences). However, this strategy tends to produce dark halo and lose contrast in some dark regions (see the "Memorial", "B-House" and "Garage" sequences).
2. The "winner-take-all" strategy is usually suitable for the scenes which contain dark regions, and it can extract enough spatial details (see the "Memorial", "B-House" and "Garage" sequences). In this situation, the dark regions are usually under-exposed in most source images, so it is more reasonable to distribute all weights to the image which has the highest local contrast. However, some bright regions may be over-exposed with this strategy. When the scene contains many objects rich in textures (see the "Forest" and "Campus" sequences), this strategy may result in low fusion quality. Furthermore, the ghost removal performance of this strategy is usually inferior to the "weighted-average" strategy. This is mainly because the moving objects often own high local contrast, especially around the object boundaries.

Multi-exposure images captured in different scenes may need different strategies. However, it is sometimes difficult to say which strategy works better since a scene may contain various contents.

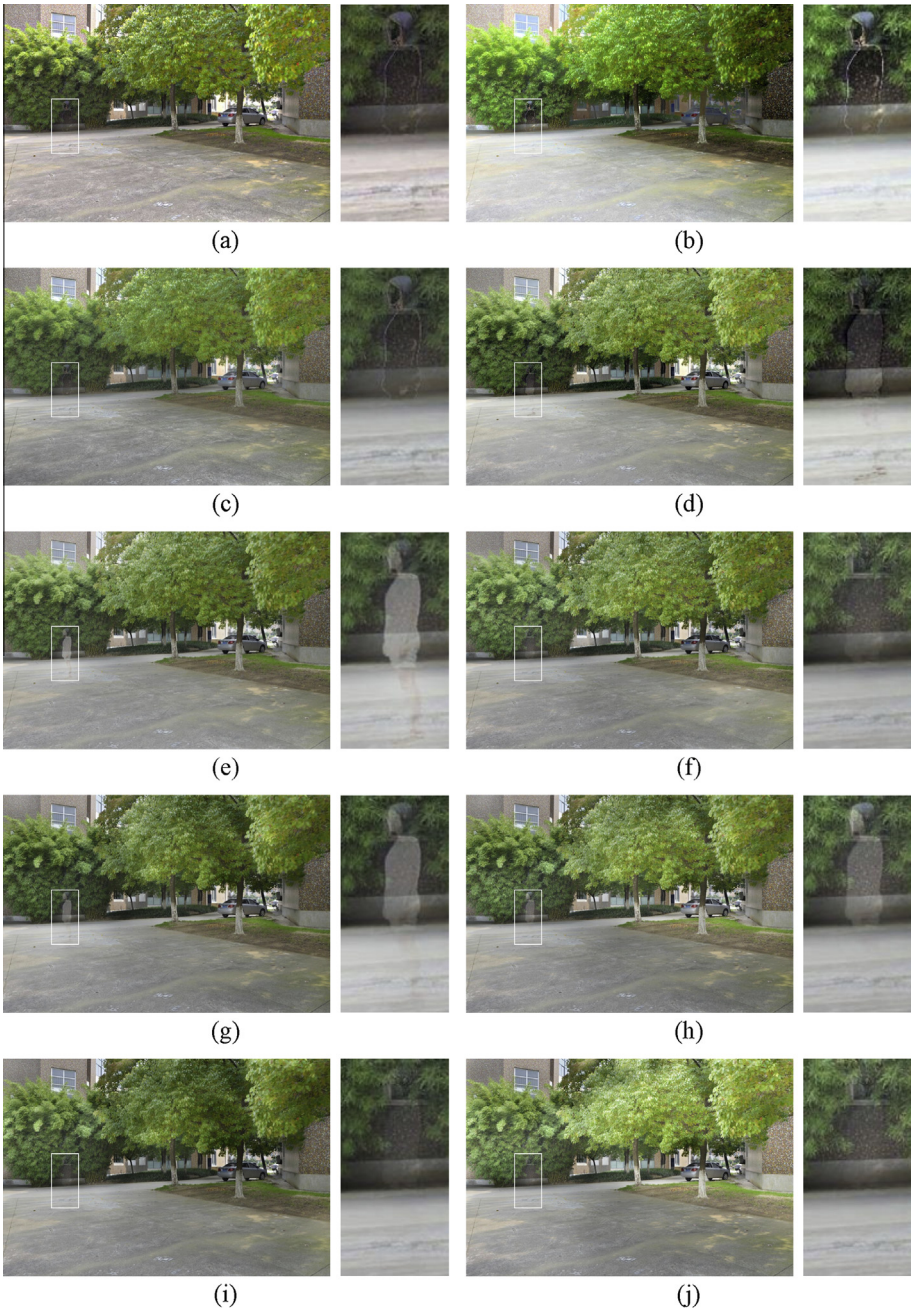

Fig. 10. Performance comparison of different methods on the "Campus" image sequence. (a) iCAM06. (b) LW. (c) Photomatix. (d) EF. (e) GRW. (f) FMMR. (g) IG-1. (h) IG-2. (i) DSIFT-1. (j) DSIFT-2.

Furthermore, there is a gap between human visual perception and photo-realistic appearance [40]. A photograph with high reality may not result in the best visual experience, and vice versa. It is practically impossible to make a result that can be preferred by all audiences. In our opinion, an ideal solution is to provide more selections and let users freely make their choices.

4.3. Objective evaluation

In this subsection, we apply two objective metrics to evaluate the performances of different methods. The first one $Q^{AB/F}$ [41] is a widely used metric in universal image fusion. $Q^{AB/F}$ is a gradient-based metric, which evaluates the extent of edge information injected into the fused image from the source images. $Q^{AB/F}$ has been adopted to evaluate the quality of multi-exposure fusion results in [17,40]. The second one is a new designed metric in this paper, aiming to evaluate the color distortion caused by the fusion process. To achieve this goal, we measure the extent of local linear correlation between the fused image and a reference HDR image. For simplicity, this metric is named $Q^{H/F}$.

4.3.1. Evaluation using $Q^{AB/F}$

In multi-exposure fusion task, most existing image fusion metrics [42] are difficult to be directly used since they are mainly designed for the situation in which only two source images exist. Please refer to [40] for more explanations. However, the fusion metric $Q^{AB/F}$ [41] is an exception because it does not need to calculate some statistics such as the covariance within source images.

Thus, this metric can be directly extended to tackle the fusion task which contains multiple source images. Moreover, the metric $Q^{AB/F}$ has been verified to be in accord better with human visual perception than many other fusion metrics [40,43]. Therefore, we employ $Q^{AB/F}$ for objective evaluation in the experiments. The extended version of $Q^{AB/F}$ for multiple source images $I_i, i = 1, 2, \ldots, N$ and the fused image I_F is defined as

$$Q^{AB/F} = \frac{\sum_{i=1}^{N}\sum_{x=1}^{H}\sum_{y=1}^{W} w^{I_i}(x,y) Q^{I_i I_F}(x,y)}{\sum_{i=1}^{N}\sum_{x=1}^{H}\sum_{y=1}^{W} w^{I_i}(x,y)}, \tag{16}$$

where $Q^{I_i I_F}(x,y) = Q_g^{I_i I_F}(x,y) Q_\alpha^{I_i I_F}(x,y)$. $Q_g^{I_i I_F}(x,y)$ and $Q_\alpha^{I_i I_F}(x,y)$ denote edge strength and orientation preservation values at pixel (x,y), respectively. $w^{I_i}(x,y)$, which is set as the gradient magnitude map of $I_i(x,y)$, indicates the weighting factor of $Q^{I_i I_F}(x,y)$. Generally, a larger $Q^{AB/F}$ value indicates a better fusion result.

The performances of different methods on the six test image sequences using $Q^{AB/F}$ are listed in Table 1, in which the highest value is shown in bold. It can be seen that the objective evaluation result is generally well consistent with the subjective visual perception. The performances of three TM-based methods are not high in most cases, which indicates these methods are sometimes not good at preserving edge information. The EF and GRW methods have close performances with other fusion methods in static sequences, but show clear disadvantages in dynamic sequences. The DSIFT-2 method obtains the highest $Q^{AB/F}$ on the "Memorial", "B-House" and "Garage" sequences, which demonstrates that the "winner-take-all" strategy tends to extract more spatial details than the "weighted-average" strategy in some scenes with dark regions. The DSIFT-1 method takes the second and first place over all the ten methods on the "Arch" and "Forest" sequences, respectively (the Photomatix method achieves the best performance on the "Arch" sequence). It is notable that the DSIFT-2 method obtains higher $Q^{AB/F}$ than the DSIFT-1 method on the "Campus" sequence. The main reason is that the result of the DSIFT-1 method is more blurred, although it has less color distortion. We can also see from Table 1 that the proposed DSIFT-based method outperforms the IG-based method with each strategy on all the six image sequences.

4.3.2. Evaluation using $Q^{H/F}$

The metric $Q^{AB/F}$ mainly focuses on the preservation of edge information, which is the most important issue in conventional image fusion tasks such as multi-focus fusion and visible-infrared fusion. However, in multi-exposure fusion, color distortion is another crucial issue. Unfortunately, most existing image fusion metrics [42] are not designed for this target. In this work, we present a simple-yet-effective metric $Q^{H/F}$ to evaluate the color fidelity of the fused image with respect to the radiance of real scene.

Table 1
Performances of different methods on the six image sequences using $Q^{AB/F}$.

Method	Memorial	B-House	Garage	Arch	Forest	Campus
iCAM06	0.5672	0.3970	0.5232	0.5388	0.4634	0.3744
LW	0.4378	0.2678	0.3623	0.3684	0.3892	0.3261
Photomatix	0.5748	0.3768	0.4523	**0.5707**	0.4379	0.3464
EF	0.5924	0.4814	0.6115	0.5426	0.3652	0.3910
GRW	0.5913	0.4832	0.6133	0.5353	0.3539	0.3507
FMMR	0.5940	0.4924	0.6121	0.5607	0.4771	0.3963
IG-1	0.5802	0.4877	0.6070	0.5598	0.4677	0.4003
IG-2	0.5818	0.4934	0.6097	0.5452	0.4641	0.4247
DSIFT-1	0.5895	0.4905	0.6094	0.5617	**0.4793**	0.4159
DSIFT-2	**0.5965**	**0.4969**	**0.6191**	0.5453	0.4697	**0.4423**

The relationship between the irradiance i of a scene and the amount of lights L captured by an imaging sensor is [44]

$$L = i \cdot \Delta t, \tag{17}$$

where Δt is the exposure time. The captured signal L is usually saved as RAW image format (has a higher dynamic with more than 8 bits) by many cameras. The ordinary 8-bit image (e.g., in JPG format) I is obtained through a nonlinear transform as

$$I = f(L), \tag{18}$$

where f is the camera response function. In HDR imaging, the irradiance values of a scene are characterized using a HDR radiance map $R(x,y)$ [19]. Therefore, with a set of exposure time $\Delta t_i, i = 1, 2, \ldots, N$, the LDR image sequence can be expressed as

$$I_i(x,y) = f(R(x,y) \cdot \Delta t_i), \quad i = 1, 2, \ldots, N. \tag{19}$$

The target of multi-exposure fusion is to obtain an everywhere well-exposed image $I_F(x,y)$. In other words, the fused image is hoped to accurately reflects an ideal amount of lights at each pixel. Considering the relationship in Eq. (17), an ideal amount of lights depends on an ideal exposure time. Since different pixels may have different ideal exposure time, we use $\Delta t(x,y)$ to denote the ideal exposure time map. Furthermore, the fused image is expected to be more faithful to the real irradiance of a scene, i.e., the measurement before applying camera response function. Thus, the fused image can be expressed as

$$I_F(x,y) = R(x,y) \cdot \Delta t(x,y). \tag{20}$$

For a local patch in the fused image, all pixels within it are assumed to have the same ideal exposure time T, namely, $I_F(x,y) = R(x,y) \cdot T$ for each pixel (x,y) within the patch. In other words, the corresponding local contents of $I_F(x,y)$ and $R(x,y)$ should be linearly correlated in the ideal case.

Based on the above analyses, we attempt to design a metric to evaluate the color fidelity of the fused image $I_F(x,y)$. Specifically, assuming that an accurate HDR radiance map $R(x,y)$ is available through some HDR reconstruction approaches, we first calculate the luminance components of $I_F(x,y)$ and $R(x,y)$ (luminance is closely related to color, and human visual system is more sensitive to luminance change than to chromatic variation [2,45]). Then, we divide $I_F(x,y)$ and $R(x,y)$ into small overlapping patches with the same size and stride. For each pair of patches, we reshape them into column vectors (denoted as $\mathbf{u}$ and $\mathbf{v}$) and calculate their *normalized inner product* $\left(\frac{\mathbf{u}^T\mathbf{v}}{\|\mathbf{u}\|_2 \cdot \|\mathbf{v}\|_2}\right)$, which indicates the extent of linear correlation between $\mathbf{u}$ and $\mathbf{v}$. A score map is obtained after handling all the patches in the same way, and the metric $Q^{H/F}$ is finally calculated as the mean value of the score map. Obviously, the range of $Q^{H/F}$ is $[0, 1]$, and a higher value indicates a better result. Moreover, the proposed metric has only two free parameters which are the patch size and the stride of sliding window. In our experiments, we set the patch size to 8×8 and the stride to one pixel. It is notable that $Q^{H/F}$ is based on the premise that a reliable HDR image is available as the reference. Since the HDR images produced by Photomatix usually own high quality, we adopt it as the reference in the evaluation.

Table 2 lists the performances of different methods on the six test image sequences using $Q^{H/F}$. In Table 2, for each sequence, the highest value among all the ten methods is shown in bold while among the seven IF-based methods is indicated with underline. Some key observations we can obtain from Table 2 are summarized as follows.

Table 2
Performances of different methods on the six image sequences using $Q^{H/F}$.

Method	Memorial	B-House	Garage	Arch	Forest	Campus
iCAM06	**0.9825**	**0.9824**	**0.9781**	0.9861	0.9396	**0.9893**
LW	0.9679	0.9603	0.9617	0.9744	0.9272	0.9853
Photomatix	0.9824	0.9753	0.9771	0.9818	0.9279	0.9843
EF	0.9788	0.9798	0.9756	0.9846	0.8849	0.9861
GRW	0.9769	0.9778	0.9744	0.9819	0.8989	0.9842
FMMR	0.9766	0.9789	0.9736	0.9850	0.9401	0.9850
IG-1	0.9788	0.9796	0.9760	0.9854	0.9381	0.9864
IG-2	0.9745	0.9768	0.9737	0.9835	0.9307	0.9852
DSIFT-1	0.9791	0.9799	0.9758	**0.9865**	**0.9403**	0.9872
DSIFT-2	0.9754	0.9772	0.9755	0.9843	0.9095	0.9842

(1) All the scores in Table 2 are very close to 1 and most are larger than 0.9. It indicates that most pairs of patches in the fused image and HDR radiance map are linearly correlated to a very high extent, which verifies the effectiveness of the "local linear correlation" assumption made above.
(2) The performances of three TM-based methods are of great differences. The iCAM06 method achieves very high scores in all the six sequences. The LW method gets the lowest scores in most cases. The performance of the Photomatix method is not very stable. This is mainly because $Q^{H/F}$ may be more suitable for some TM operators while less for others, depending on the specific scheme of a TM operator.
(3) The "weighted-average" strategy universally outperforms the "winner-take-all" strategy in terms of color fidelity. For all the six sequences, the IG-1 and DSIFT-1 methods obtain higher scores than the IG-2 and DSIFT-2 methods, respectively. It is notable that the DSIFT-2 method gets a very low score in the "Forest" sequence, which corresponds with the subjective visual perception and partly verifies the ability of $Q^{H/F}$ for color distortion evaluation.
(4) The performances of the EF and GRW methods are relatively lower in dynamic fusion than static fusion. For example, the differences between the EF method and the DSIFT-1 method are clearly larger in the last three sequences. This is mainly due to the lack of deghosting ability of the EF and GRW methods. In particular, the difference is the most obvious in the "Forest" sequence since the moving object covers a large proportion of pixels in that sequence.
(5) The DSIFT-1 method generally obtains the highest scores among all the seven fusion methods. Although the scores of the IG-1 method the DSIFT-1 method are very close in three static fusion examples, the DSIFT-1 method exhibits clear advantages over the IG-1 method in three dynamic image sequences.

In summary, the objective evaluation using $Q^{AB/F}$ and $Q^{H/F}$ is well consistent with the subjective visual perception. The DSIFT-based method owns advantages over the IG-based method, especially for ghost removal in dynamic scenes. The "weighted-average" strategy can preserve color information better than the "winner-take-all" strategy, but the "winner-take-all" strategy can usually extract more spatial details and sometimes likely to provide better visual experience. As mentioned in Section 4.2, a photograph with higher reality is not equivalent to a better visual experience, and we believe the choice of strategy is scene dependent as well as user dependent.

4.4. Impact of free parameter

As mentioned in Section 4.1, the scale factor used for dense SIFT calculation [25] in Eq. (1) is fixed at 16 in the above experiments. In this subsection, we apply $Q^{AB/F}$ and $Q^{H/F}$ to quantitatively study the impact of this free parameter. To achieve this goal, we set the scale factor to 8, 12, 16, 20, 24, 28 and 32, respectively. The two versions of the DSIFT-based method are both tested, i.e., DSIFT-1 and DSIFT-2. For each version with a certain scale factor, the average metric value of the six image sequences is employed for evaluation.

Fig. 11 shows the objective performances of the DSIFT-1 and DSIFT-2 methods with different scale factors. On one hand, it can be seen from Fig. 11(a) that for each of the two versions, the best performance on $Q^{AB/F}$ is obtained when the scale factor is equal to 16. On the other hand, for metric $Q^{H/F}$ shown in Fig. 11(b), the scale factor has little effect on the performance of the DSIFT-1 method, and the performance of the DSIFT-2 method slightly decreases when the scale factor increases from 8 to 32. Overall, it is a reasonable choice to set the scale factor to 16. Furthermore, the variation range of either $Q^{AB/F}$ or $Q^{H/F}$ is very small when the scale factor increases from 8 to 32 for both DSIFT-1 and DSIFT-2, which means that the proposed method is not sensitive to the change of this parameter.

4.5. Computational efficiency

In this subsection, we compare the computational efficiency of the above multi-exposure fusion methods. For either the IG or

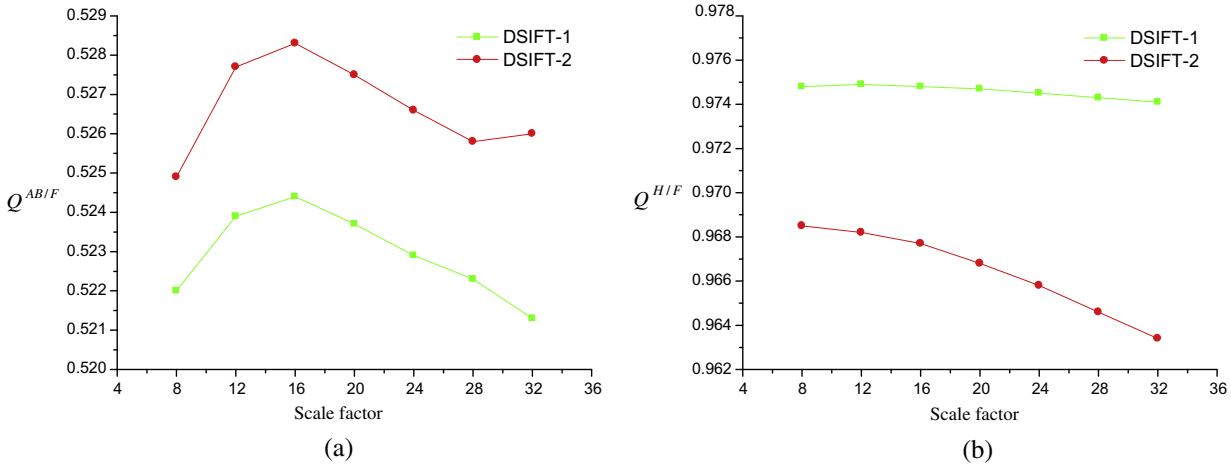

Fig. 11. Objective performances of the proposed fusion method with different scale factors. (a) $Q^{AB/F}$. (b) $Q^{H/F}$.

Table 3
Computational efficiency of different fusion methods on the six image sequences (unit: second).

Sequence	Size	EF	GRW	FMMR	IG	DSIFT	DSIFT ↑
Memorial	464 × 696 × 16	4.52	1.26	9.87	6.85	13.2	4.76
B-House	1025 × 769 × 9	6.31	2.07	12.7	10.5	19.8	6.25
Garage	348 × 222 × 6	0.43	0.13	0.82	0.74	1.33	0.56
Arch	669 × 1024 × 5	3.30	1.21	10.1	7.02	23.8	6.28
Forest	1024 × 683 × 4	2.76	1.07	18.9	5.42	17.9	4.61
Campus	1200 × 800 × 4	3.68	1.46	27.3	7.17	24.2	6.36

DSIFT based method, the selection of strategy has almost no influence on the running time, so just one measure is given here. All the methods are implemented with MATLAB on a computer with a 3.0 GHz CPU. The testing results are listed in Table 3. It can be seen that the GRW and EF methods are faster than other methods, but they cannot tackle the fusion task in dynamic scenes. The DSIFT method is more time-consuming than the IG method as well as the FMMR method in most cases. The speed of our method can be accelerated by operating all the weight maps with half of the original spatial resolution, namely, the source images are first down-sampled and the obtained weight maps are then up-sampled for merging. We experimentally verify that this acceleration scheme has little impact on the final fusion quality. Table 4 lists the objective performances of the proposed method and its accelerated version DSIFT ↑ using $Q^{AB/F}$. (Since this scheme mainly affects the extraction of local details, only $Q^{AB/F}$ is used here.) We can see that the performance of the DSIFT ↑ method is still better than the IG method in most cases, especially for dynamic fusion. The running time of DSIFT ↑ on the six image sequences are listed in the last column in Table 3. It can be seen that the efficiency of DSIFT ↑ is much improved and even higher than that of the IG method. Moreover, with a more efficient programming language, the efficiency of our method can be further improved.

Table 4
Objective performances of the proposed method and its accelerated version using $Q^{AB/F}$.

Method	Memorial	B-House	Garage	Arch	Forest	Campus
IG-1	0.5802	0.4877	0.6070	0.5598	0.4677	0.4003
DSIFT-1	0.5895	0.4905	0.6094	0.5617	0.4793	0.4159
DSIFT-1 ↑	0.5824	0.4883	0.6076	0.5615	0.4751	0.4112
IG-2	0.5818	0.4934	0.6097	0.5452	0.4641	0.4247
DSIFT-2	0.5965	0.4969	0.6191	0.5453	0.4697	0.4423
DSIFT-2 ↑	0.5815	0.4928	0.6136	0.5465	0.4693	0.4352

4.6. Other applications

In [15], the authors employ their IG-based exposure fusion method to the application of flash and no-flash photography [46]. The proposed method can also be used for this application. Fig. 12 shows a flash/no-flash image pair and their combination results of different fusion methods. The flash image shown in Fig. 12(a) captures more details but suffers from hot-spot artifacts, while the no-flash image shown in Fig. 12(b) reveals ambient illumination. It can be seen that the DSIFT-2 method can obtain the highest visual quality among all the methods (see the largest red leaf, i.e., the hot-spot region in Fig. 12(c)–(i)). The fused result shown in Fig. 12(i) is completely free of hot spots and all the details are well preserved.

Fig. 12. Performance comparison of different fusion methods on a pair of flash and no-flash images. (a) Flash image. (b) No-flash image. (c) EF. (d) GRW. (e) FMMR. (f) IG-1. (g) IG-2. (h) DSIFT-1. (i) DSIFT-2.

4.7. Further discussions

In [19], the authors concluded that "*there is no single best method and the selection of an approach depends on the user's goal*". The proposed method follows the framework of weighted sum based ghost-free exposure fusion methods [15–17], so it still has some common limitations owned by these approaches.

The first issue is the request for a minimum number of input LDR images [19]. When the moving objects do not appear in a small percentage of source images at one location, the proposed method may not work well. For instance, considering a scene which is a square with many walking people. Then, for a certain pixel location in most of exposures, it may be covered by different people or the same person but with a motion. In this situation, ghosting artifacts tend to be produced. We experimentally find that the minimum number of LDR images in our method is about four to ensure a generally satisfactory deghosting performance. The reference image based methods such as [23] can sometimes achieve better results when the number of input images is smaller, but they often introduce some new artifacts such as the seams at the boundaries of ghost regions [19]. In addition, the quality of fusion result depends greatly on the selection of reference image as well as the specific scene. Please refer to the Supplementary material for more results and discussions about this issue.

Another defect is the moving background such as the wind-blown leaves or waves, which has been mentioned before in the "Campus" image sequence. In this situation, the fusion results usually suffer from blurry artifacts. The artifacts caused by camera shaking or image mis-alignment are similar to those caused by the moving background. The reference image based approaches can often have a better performance in this situation, but it is still scene dependent. It is notable that the proposed DSIFT-based ghost removal measure can be easily modified to be applicable to the reference image based fusion scheme, just as the modification done by Zhang and Cham from [15] to [23].

5. Conclusion and future work

In this paper, we present a new multi-exposure fusion method with dense SIFT. In our algorithm, the unnormalized dense SIFT descriptor is employed as the activity level measurement to extract local details, and the normalized dense SIFT descriptor is used for the removal of ghosting artifacts when the captured scene is dynamic. Experimental results shows that the proposed method can outperform many representative tone mapping and image fusion based methods in terms of both visual quality and objective evaluation. Moreover, two weight distribution strategies for local contrast extraction, namely, "weighted-average" and "winner-take-all" are studied in this paper. Some discussions based on the experimental results are given to help users make good choices. In the future, by further taking the advantages of dense SIFT such as the ability of local feature matching, we plan to develop more effective ghost-free fusion scheme to overcome the limitations discussed above.

Acknowledgments

The authors would like to thank the editors and anonymous reviewers for their detailed review, valuable comments and constructive suggestions. The authors would also like to thank Dr. Rui Shen for providing us with the executable file of the GRW method [6]. This work is supported by the National Natural Science Foundation of China (No. 61472393) and the National Science and Technology Projects (No. 2012GB102007).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jvcir.2015.06.021.

References

[1] P.E. Debevec, J. Malik, Recovering high dynamic range radiance maps from photographs, in: Proc. ACM SIGGRAPH, 1997, pp. 369–378.
[2] E. Reinhard, G. Ward, S. Pattanaik, P. Debevec, High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting, Morgan Kaufman Publishers, 2005.
[3] E. Reinhard, M. Stark, P. Shirley, J. Ferwerda, Photographic tone reproduction for digital images, in: Proc. ACM SIGGRAPH, 2002, pp. 267–276.
[4] J. Kuang, G.M. Johnson, M.D. Fairchild, iCAM06: a refined image appearance model for HDR image rendering, J. Vis. Comun. Image Represent. 18 (5) (2007) 406–414.
[5] Q. Shan, J. Jia, M.S. Brown, Globally optimized linear windowed tone mapping, IEEE Trans. Vis. Comput. Graph. 16 (4) (2010) 663–675.
[6] R. Shen, I. Cheng, J. Shi, A. Basu, Generalized random walks for fusion of multi-exposure images, IEEE Trans. Image Process. 20 (12) (2011) 3634–3646.
[7] T. Stathaki, Image Fusion: Algorithms and Applications, Academic Press, 2008.
[8] A. Goshtasby, S. Nikolov, Image fusion: advances in the state of the art, Inform. Fusion 8 (2) (2007) 114–118.
[9] G. Piella, Image fusion for enhanced visualization: a variational approach, Int. J. Comput. Vis. 83 (1) (2009) 1–11.
[10] S. Li, X. Kang, Image fusion with guided filtering, IEEE Trans. Image Process. 22 (7) (2013) 2864–2875.
[11] A. Goshtasby, Fusion of multi-exposure images, Image Vision Comput. 23 (6) (2005) 611–618.
[12] T. Mertens, J. Kautz, F.V. Reeth, Exposure fusion, in: Proc. Pacific Graphics, 2007, pp. 382–390.
[13] B. Gu, W. Li, J. Wong, M. Zhu, M. Wang, Gradient field multi-exposure images fusion for high dynamic range image visualization, J. Vis. Comun. Image Represent. 23 (4) (2012) 604–610.
[14] M. Song, D. Tao, C. Chen, J. Bu, J. Luo, C. Zhang, Probabilistic exposure fusion, IEEE Trans. Image Process. 21 (1) (2012) 341–357.
[15] W. Zhang, W.-K. Cham, Gradient-directed multi-exposure composition, IEEE Trans. Image Process. 21 (4) (2012) 2318–2323.
[16] J. An, S. Lee, J. Kuk, N. Cho, A multi-exposure image fusion algorithm without ghost effect, in: Proc. IEEE ICASSP, 2011, pp. 1565–1568.
[17] S. Li, X. Kang, Fast multi-exposure image fusion with median filter and recursive filter, IEEE Trans. Consum. Electron. 58 (2) (2012) 626–632.
[18] E.S.L. Gastal, M.M. Oliveira, Domain transform for edge-aware image and video processing, ACM Trans. Graph. 30 (4) (2011) 69.
[19] A. Srikantha, D. Sidibe, Ghost detection and removal for high dynamic range images: recent advances, Signal Process.: Image Commun. 27 (6) (2012) 650–662.
[20] P. Sen, N. Kalantari, M. Yaesoubi, S. Darabi, D. Goldman, E. Shechtman, Robust patch-based hdr reconstruction of dynamic scenes, ACM Trans. Graph. 31 (6) (2012) 203.
[21] M. Granados, K. Kim, J. Tompkin, C. Theobalt, Automatic noise modeling for ghost-free hdr reconstruction, ACM Trans. Graph. 32 (6) (2013) 201.
[22] K. Hadziabdic, J. Telalovic, Mantiuk, Comparison of deghosting algorithms for multi-exposure high dynamic range imaging, in: Proc. ACM Spring Conference on Computer Graphics, 2013, pp. 21–28.
[23] W. Zhang, W.-K. Cham, Reference-guided exposure fusion in dynamic scenes, J. Vis. Commun. Image Represent. 23 (3) (2012) 467–475.
[24] C. Wang, C. Tu, An exposure fusion approach without ghost for dynamic scenes, in: Proc. CISP, 2013, pp. 904–909.
[25] C. Liu, J. Yuen, A. Torralba, Sift flow: dense correspondence across scenes and its applications, IEEE Trans. Pattern Anal. Mach. Intell. 33 (5) (2011) 978–994.
[26] Y. Liu, S. Liu, Z. Wang, Multi-focus image fusion with dense SIFT, Inform. Fusion 23 (1) (2015) 139–155.
[27] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[28] H. Li, B. Manjunath, S. Mitra, Multisensor image fusion using the wavelet transform, Graph. Models Image Process. 57 (3) (1995) 235–245.
[29] A. Tomaszewska, R. Mantiuk, Image registration for multi-exposure high dynamic range image acquisition, in: International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG), 2007, pp. 49–56.
[30] G. Ward, Fast, robust image registration for compositing high dynamic range photographs from handheld exposures, J. Graph. Tools 8 (2) (2003) 17–30.
[31] R. Gonzalez, R. Woods, S. Eddins, Digital Image Processing using MATLAB, second ed., Gatesmark Publishing, 2009.
[32] S. Paris, F. Durand, A fast approximation of the bilateral filter using a signal processing approach, in: Proc. ECCV, 2006, pp. 568–580.
[33] K. He, J. Sun, X. Tang, Guided image filtering, IEEE Trans. Pattern Anal. Mach. Intell. 35 (6) (2013) 1397–1409.
[34] T. Mertens. <http://research.edm.uhasselt.be/~tmertens/>.
[35] X. Kang. <http://xudongkang.weebly.com/index.html>.
[36] Photomatix. <http://www.hdrsoft.com/>.

[37] K. Hadziabdic, J. Telalovic, Mantiuk, Expert evaluation of deghosting algorithms for multi-exposure high dynamic range imaging, in: Proc. Second International Conference and SME Workshop on HDR imaging, 2014.
[38] J. Kuang. <http://www.cis.rit.edu/research/mcsl2/icam06/>.
[39] Q. Shan. <http://homes.cs.washington.edu/~shanqi/>.
[40] R. Shen, I. Cheng, A. Basu, Qoe-based multi-exposure fusion in hierarchical multivariate gaussian crf, IEEE Trans. Image Process. 22 (6) (2013) 2469–2478.
[41] C.S. Xydeas, V.S. Petrovic, Objective image fusion performance measure, Electron. Lett. 36 (4) (2000) 308–309.
[42] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, W. Wu, Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: a comparative study, IEEE Trans. Pattern Anal. Mach. Intell. 34 (1) (2012) 94–109.
[43] V. Petrovic, Subjective tests for image fusion evaluation and objective metric validation, Inform. Fusion 8 (2) (2007) 208–216.
[44] T. Jinno, M. Okuda, Multiple exposure fusion for high dynamic range image acquisition, IEEE Trans. Image Process. 21 (1) (2012) 358–365.
[45] WorkWithColor. <www.workwithcolor.com/color-luminance-2233.htm>.
[46] G. Petschnigg, M. Agrawala, H. Hoppe, R. Szeliski, M. Cohen, K. Toyama, Digital photography with flash and no-flash image pairs, ACM Trans. Graph. 23 (3) (2004) 664–672.