Information Fusion 23 (2015) 139–155

Contents lists available at ScienceDirect

Information Fusion

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

Multi-focus image fusion with dense SIFT

Yu Liu [a], Shuping 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 2014
Received in revised form 26 February 2014
Accepted 21 May 2014
Available online 10 June 2014

Keywords:
Multi-focus image fusion
Dense SIFT
Feature space transform
Activity level measurement
Local feature matching

ABSTRACT

Multi-focus image fusion technique is an important approach to obtain a composite image with all objects in focus. The key point of multi-focus image fusion is to develop an effective activity level measurement to evaluate the clarity of source images. This paper proposes a novel image fusion method for multi-focus images with dense scale invariant feature transform (SIFT). The main novelty of this work is that it shows the great potential of image local features such as the dense SIFT used for image fusion. Particularly, the local feature descriptor can not only be employed as the activity level measurement, but also be used to match the mis-registered pixels between multiple source images to improve the quality of the fused image. In our algorithm, via the sliding window technique, the dense SIFT descriptor is first used to measure the activity level of source image patches to obtain an initial decision map, and then the decision map is refined with feature matching and local focus measure comparison. Experimental results demonstrate that the proposed method can be competitive with or even outperform the state-of-the-art fusion methods in terms of both subjective visual perception and objective evaluation metrics.

1. Introduction

Due to the finite depth-of-field (DOF) of optical lenses, it is usually difficult for cameras to capture an image in which all the objects are in focus. Generally, only the objects within the DOF appear sharp in the captured photograph while other objects tend to be blurred [1]. A popular low-cost approach to solve this problem is multi-focus image fusion technique, which aims at obtaining an all-in-focus image by integrating the complementary information from multiple images taken with diverse focal lengths. The composite image also known as the fused image will be more suitable for machine/human perception. Multi-focus image fusion has become a representative topic since many algorithms can be naturally extended to other image fusion applications such as remote sensing and medical imaging. During the past two decades, a great variety of image fusion algorithms have been proposed. Conventionally, these algorithms are classified into transform domain methods and spatial domain methods by most publications in the literature [2,3]. Since many new fusion algorithms have been presented recently, we would like to make a more detailed classification to group the existing fusion methods into four categories: multi-scale transform methods, feature space transform methods, spatial domain methods, and pulse coupled neural networks (PCNN) methods.

The multi-scale transform fusion methods [4] have a relative long history in which pyramid decomposition [5,6], wavelet transform [7,8] and multi-scale geometric analysis [9–11] are sequentially applied into image fusion. Specifically, these methods includes Laplacian pyramid (LP) [5], gradient pyramid (GP) [6], discrete wavelet transform (DWT) [7], dual tree complex wavelet transform (DTCWT) [8], curvelet transform (CVT) [9] and non-subsampled contourlet transform (NSCT) [10,11], etc. Li et al. [12] made a comparative study of different multi-scale transform fusion methods. They found that the NSCT-based method can generally achieve the best results. Recently, Zhao et al. [13] proposed a new multi-scale transform fusion method based on neighbor distance (ND), which can generally achieve better results than the NSCT-based method. All these multi-scale transform methods share a "decomposition–fusion–reconstruction" framework, i.e., decompose the source images into a multi-scale domain, fuse the transformed coefficients by a given rule and reconstruct the fused image with the fused coefficients. In addition to the selection of transform domain, the fusion rule in high- or low-frequency domain also plays a crucial role in these methods.

In the past few years, the feature space transform fusion methods have become a popular topic in current image fusion research. Unlike the traditional multi-scale transform methods, this category

* Corresponding author at: Department of Automation, University of Science and Technology of China, Hefei 230026, China. Tel.: +86 551 63600634.
E-mail addresses: liuyu1@mail.ustc.edu.cn (Y. Liu), fengya@mail.ustc.edu.cn (S. Liu), zfwang@ustc.edu.cn (Z. Wang).

http://dx.doi.org/10.1016/j.inffus.2014.05.004
1566-2535/

of methods inclines to measure the clarity (i.e., activity level) of source images in a single-scale feature space. Some representative examples are the independent component analysis (ICA)-based fusion method [14], the sparse representation (SR)-based method [15], the higher order singular value decomposition (HOSVD)-based method [16], the robust principal component analysis (RPCA)-based method [17], etc. The key point of these methods is to find a reliable feature space which can reflect the activity level of image patches. In addition, these methods apply the sliding window technique to make the fusion process shift-invariant, which is of great importance for image fusion. In general, this kind of methods can achieve better results when compared with traditional multi-scale transform methods.

The earliest spatial domain fusion method calculates the average of source images pixel-by-pixel, which usually causes undesirable artifacts. In recent years, some block-based methods [18,19] and region-based methods [20,21] have been proposed. The basic principle of these methods is to select image blocks or segmented regions from source images with some focus measure metrics [22], such as image variance, spatial frequency, etc. However, as an image block may simultaneously contain both clear and blurry areas, the fused images of block-based methods usually suffer from blocking effect even though the block size can be adaptively determined. For the region-based methods, it is hard to stably obtain fused result with high quality since image segmentation is also a difficult task. During the past three years, many novel pixel-based spatial domain fusion algorithms have been introduced. These methods can usually extract enough details from the source images and preserve the spatial consistency of the fused image well. In particular, the most recent image matting (IM)-based method [23] and guided filtering (GF)-based method [24] can obtain the state-of-the-art results with high computational efficiency.

The last category is the PCNN fusion methods. PCNN is a biological neural network developed by Eckhorn et al. [25] based on cat visual cortex, and it has been successfully employed in many image processing applications including image fusion. For image fusion task, PCNN can be used either in multi-scale transform domain [11,26] or directly in spatial domain [20,27]. The most important advantage of PCNN fusion methods is that their information processing mode is in accord with human visual system. However, there are a large number of free parameters in PCNN model and the fusion performance is often sensitive to them. It should be noticed that there is no definite boundary between the above four categories of fusion methods for some fusion methods [11,20] can be viewed as mixed ones. In addition to the PCNN model, both the feature space transform fusion strategies and the spatial domain fusion strategies can also be applied under the multi-scale transform framework.

Although some of the recent fusion algorithms can achieve good results for multi-focus images, there is still large room for improvement based upon the following three considerations. First, the effectiveness of the feature space transform fusion methods has been demonstrated, but the related study is just at the preliminary stage. It is necessary to develop some more effective feature spaces for the activity level measurement of source images. Second, few researchers have addressed the problem of fusing multi-focus images in dynamic scene that contains moving objects or camera movements during different captures [23]. Furthermore, due to the different imaging parameters for multiple source images, the location of object edges in different source images cannot be exactly the same. As we know, the merging effects of either the object motion regions or the object edges have a great impact on the quality of final fused images, but most of the existing fusion methods cannot work well in these situations. Third, the computational cost is high for some fusion methods, especially for some feature space transform methods.

In this paper, we propose a novel feature space transform fusion method for multi-focus images with dense scale invariant feature transform (DSIFT). In our method, via the sliding window technique, the dense SIFT descriptor is first used to measure the activity level of source image patches to obtain an initial decision map, and then the decision map is refined with feature matching and local focus measure comparison. Thus, the proposed algorithm also has some distinct characteristics of spatial domain fusion methods. To the best of our knowledge, image local feature descriptors have not been directly applied to image fusion so far. The most significant contribution of this paper is it indicates that some image local features such as the dense SIFT can construct an effective feature space for image fusion. In particular, the local feature descriptors cannot only be employed as the activity level measurement, but also be used to match the mis-registered pixels between multiple source images to improve the fused quality of both object motion regions and object edges. Experimental results on twelve pairs of multi-focus images demonstrate that the proposed fusion method can be competitive with or even outperform some state-of-the-art fusion methods in terms of both subjective visual perception and six objective assessment metrics.

The rest of this paper is organized as follows. In Section 2, the dense SIFT and its relationship with our fusion algorithm are presented. In Section 3, the proposed multi-focus image fusion algorithm using dense SIFT is introduced in detail. Section 4 presents the experimental results and discussions. Finally, Section 5 concludes the paper and puts forward the future work.

2. Dense SIFT for image fusion

Image local features and descriptors play a pivotal role in various computer vision applications, such as image registration and object recognition. Lowe [28] proposed the most popular SIFT algorithm including both feature detection and description stages. The rotation and scale invariant SIFT descriptor is generated by characterizing the local gradient information around a corresponding detected interest point. Therefore, for image fusion task, the SIFT descriptor may contain underlying activity level information of source images, which generates the original motivation of this work. However, for a whole image, only the interest points own such descriptors, i.e., the original SIFT in [28] is a sparse feature representation method for an image. As dense feature representation method is preferred in image fusion, the original SIFT descriptor cannot be directly used.

Recently, Liu et al. [29,30] introduced the dense SIFT descriptor for image registration and object recognition. In dense SIFT, there is no feature detection stage while local feature descriptors are extracted at every pixel to obtain a pixel-wise SIFT image [30]. Specifically, for each pixel in an image, its neighborhood patch with a certain size like 48×48 is first divided into a 4×4 cell array. In each cell, an orientation histogram with 8 bins is used to quantize the gradient information. The histogram is constructed by accumulating the gradient orientations of all the pixels in a cell weighted by their gradient magnitudes. Thus, a $4 \times 4 \times 8 = 128$ dimensional feature vector can be formed as the feature descriptor. Finally, the original descriptors are usually required to be normalized with the method in [28] for robust feature matching. The most important parameter in dense SIFT is the size of neighborhood patch, which denotes the scale factor of image features. It should be noted that the dense SIFT descriptor is not rotation and scale invariant since all the pixels in an image use a fixed-size patch as the neighborhood. Fortunately, the multiple source images are pre-registered in image fusion task, thereby neither rotation nor scale invariance is required.

In this work, the unnormalized dense SIFT descriptor is used to construct a space for feature extraction. Particularly, the sum of all

the elements (each element is always positive) in a descriptor vector is employed as activity level measurement of source images. To confirm the feasibility of this idea, we conduct a simple experiment with four sets of synthetic images. Fig. 1(a) shows four test images of size 512×512. For each test image, it is blurred by the Gaussian filter with different standard deviations from 0.5 to 4 with a sampling interval of 0.1. For each blurred image, we first calculate the dense SIFT descriptors at every pixel, and the size of neighborhood patch is set to 48×48 here, i.e., the scale factor is 48. In this work, the code in website [31] provided by C. Liu (the first author of [29,30]) is used for dense SIFT calculation. In order to ensure that the boundary pixels in an image can also own descriptors, the input image is extended with appropriate number of "zero-valued" pixels before dense SIFT calculation. Then, the sum of all the elements in a descriptor vector is calculated, so a sum map with the same size of original image can be obtained. Finally, the average value of all the coefficients in the sum map is used as the output to measure the overall clarity of the input image. Fig. 1(b) shows the relation curves of standard deviation and overall clarity of the four test images. It can be seen that for each of the four images, the clarity measurement clearly decreases when the standard deviation increases before approaching to 3. The clarity measurement tends to be stable when the standard deviation is larger than 3, which is mainly because that the Gaussian filter will get close to a mean filter when its standard deviation becomes large enough. Furthermore, the curves of "mountain" and "baboon" are steeper than those of "lena" and "boat" for the reason that "mountain" and "baboon" contain more spatial details while there are many flat regions in "lena" and "boat". On the whole, this simple test shows the potential of the unnormalized dense SIFT descriptor used as an activity level measurement for image fusion.

In addition to the unnormalized descriptor, the normalized descriptor is also employed in the proposed fusion approach. As mentioned before, the fusion problem of either object motion regions or object edges is not well tackled in most previous work. The main reason is that the locations of moving objects and object edges in different source images cannot be exactly the same, while most traditional methods are based on the assumption that different source images are accurately pre-registered. As image local feature descriptor is capable of finding corresponding pixels between two images, it may has potential advantages over traditional activity level measurements. In this work, the normalized dense SIFT descriptor is used to match the mis-registered pixels between multiple source images, which can improve the fused quality of both object motion regions and object edges. The detailed strategy will be presented in the next section.

3. Detailed fusion scheme

The schematic diagram of the proposed fusion algorithm is illustrated in Fig. 2. For simplicity, we only consider the task of fusing two source images at first, and the generalized algorithm to process more than two images will be presented at the end of this Section. Moreover, the source images are also assumed to be pre-registered in this work. From Fig. 2, we can see that there are four main steps of the proposed fusion method. First, the 128D SIFT image of each source image is calculated, and the activity level map which contains focus information is formed by accumulating all the elements of an unnormalized SIFT descriptor. Simultaneously, the normalized SIFT image is obtained with the normalization rule in [28]. Then, the focus information in two activity level maps are combined together to generate an initial decision map, which consists of the definite focused regions of the first source image, the definite focused regions of the second source image and indefinite regions. Next, the initial decision map is refined with feature matching and local focus measure comparison. Finally, the fused image is obtained with the final decision map.

In the proposed algorithm, color source images are first converted to gray images by $I = 0.299r + 0.587g + 0.114b$ [32], where r, g and b denote the red, green and blue channel respectively. In Fig. 2, the SIFT images are visualized by projecting the 128D descriptor into a 3D RGB color space using the approach in [30]. In a projected SIFT image, pixels with similar colors share similar local structures. The details of calculating dense SIFT have been introduced in Section 2. The next three steps are denoted as P1, P2, and P3 in Fig. 2. In the next three subsections, we will provide

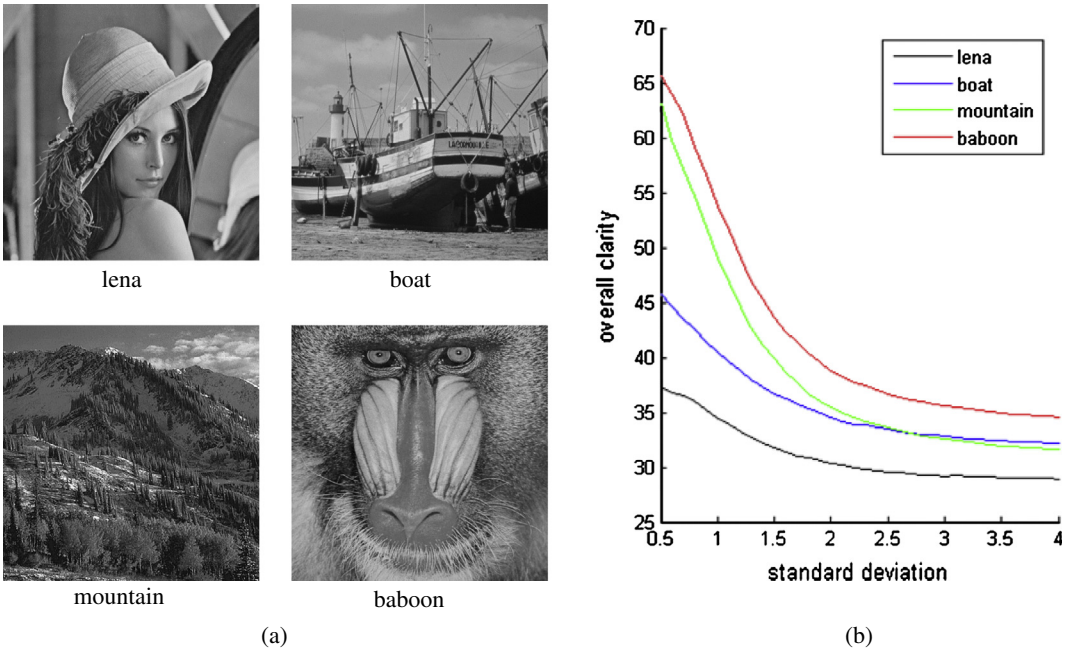


Fig. 1. The feasibility of dense SIFT descriptor used as activity level measurement. (a) Test images and (b) relation curves.

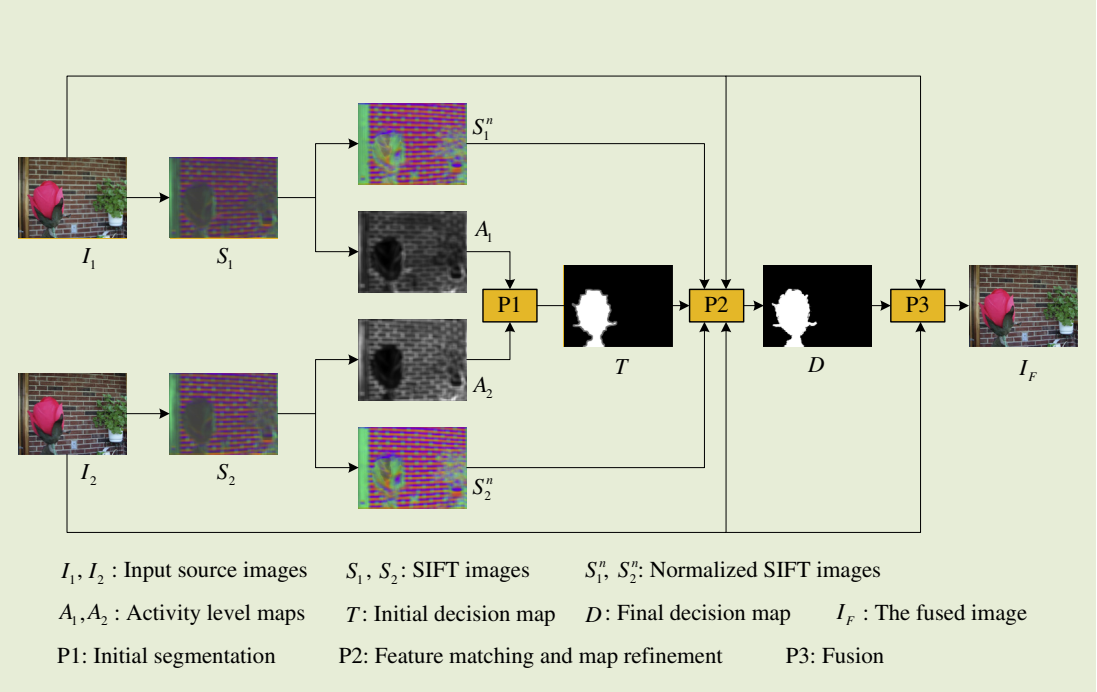


Fig. 2. Schematic diagram of the proposed fusion algorithm.

detailed descriptions for them respectively. Furthermore, to have a better observation of the example flower images used in Fig. 2, we exhibit the results of each step separately. Fig. 3 shows the source images, activity level maps and normalized SIFT images.

3.1. Initial segmentation

With the activity level maps obtained from dense SIFT descriptors, the next target is to generate an initial decision map, which consists of the definite focused regions of the first source image (the definite defocused regions of the second source image), the definite focused regions of the second source image (the definite defocused regions of the first source image) and indefinite regions. Unlike conventional pixel-based activity level measurements, we use the sum of the coefficients within a local patch in the activity level map to evaluate the clarity of the corresponding image patch. Moreover, the sliding window technique is applied to make the fusion approach shift-invariant. Specifically, the process of generating initial decision map takes the following three substeps.

(i) Apply the sliding window technique to divide the activity level maps A_1 and A_2 into patches of size $n \times n$ from upper left to lower right with a step length of one pixel, as shown in Fig. 4. Thus, if the spatial resolution of each source image

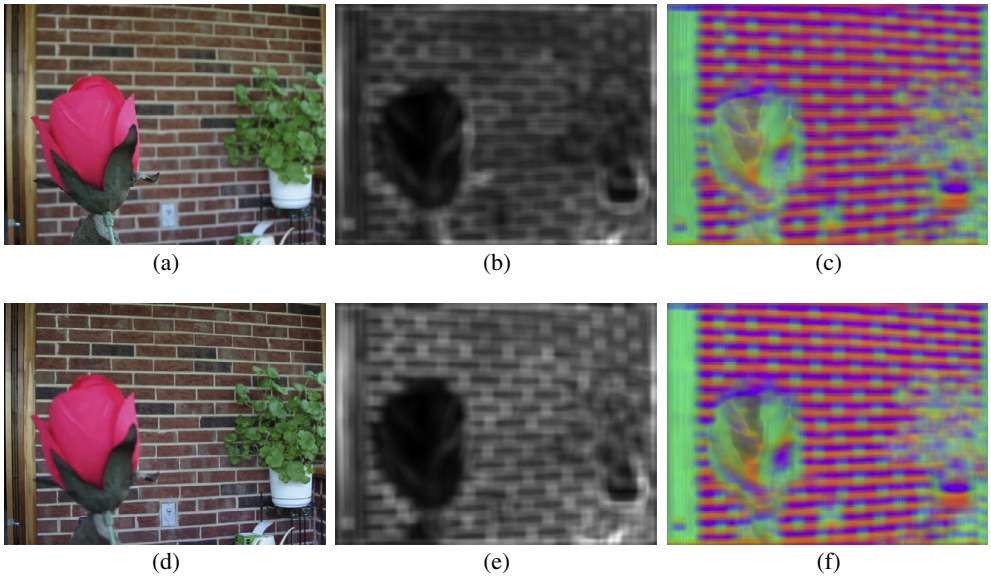


Fig. 3. Dense SIFT calculation of flower images. (a) Near focused source image, (b) the activity level map of (a), (c) the normalized SIFT image of (a), (d) far focused source image, (e) the activity level map of (d), and (f) the normalized SIFT image of (d).

is $H \times W$, there will be $Q = (H - n + 1) \times (W - n + 1)$ patches denoted as $\{p_1^i\}_{i=1}^Q$ and $\{p_2^i\}_{i=1}^Q$ in A_1 and A_2, respectively.

(ii) Create two new "zero-valued" maps M_1 and M_2 to store **focus scores**, which are used to measure the activity level of each pixel in the two source images. As shown in Fig. 4, for each corresponding pair of patches p_1^i and p_2^i, calculate the sum of all the coefficients in each of them. Let s_1^i and s_2^i denote the sum values, respectively. If $s_1^i > s_2^i$, all the coefficients within the corresponding patch in M_1 plus one (the position of this patch in M_1 is same as that of p_1^i in A_1), and vice versa. After iterating the above process for all corresponding patches in $\{p_1^i\}_{i=1}^Q$ and $\{p_2^i\}_{i=1}^Q$, the two score maps M_1 and M_2 are constructed. For an arbitrary position (x,y), $M_1(x,y) + M_2(x,y)$ records the total times of the above comparison between s_1^i and s_2^i occurring at (x,y). It is worthwhile to note that the sum of $M_1(x,y)$ and $M_2(x,y)$ satisfies

$$M_1(x,y) + M_2(x,y) \leqslant n^2, \tag{1}$$

where n is the side length of sliding patches as defined before. As the sliding window technique is applied, the position (x,y) may be simultaneously included in several overlapped patches. For an inner pixel, it is easy to find that $M_1(x,y) + M_2(x,y)$ always equals to n^2. While for a pixel near the image border, the sum is less than n^2.

(iii) Based on the obtained focus score maps $M_1(x,y)$ and $M_2(x,y)$, every pixel in each source image is classified into three categories: focused pixel, defocused pixel and uncertain pixel. The classification rule for a pixel (x,y) in the first source image I_1 is

$$I_1(x,y) \text{ is a/an} \begin{cases} \text{focused pixel}, & \text{if } M_2(x,y) = 0 \\ \text{defocused pixel}, & \text{if } M_1(x,y) = 0 \\ \text{uncertain pixel}, & \text{otherwise} \end{cases}. \tag{2}$$

Correspondingly, the classification rule for a pixel (x,y) in the second source image I_2 is

$$I_2(x,y) \text{ is a/an} \begin{cases} \text{focused pixel}, & \text{if } M_1(x,y) = 0 \\ \text{defocused pixel}, & \text{if } M_2(x,y) = 0 \\ \text{uncertain pixel}, & \text{otherwise} \end{cases}. \tag{3}$$

The above two rules mean that a pixel in a source image will be viewed as a focused/defocused pixel only when it wins/loses all the above comparisons between s_1^i and s_2^i occurring at its position.

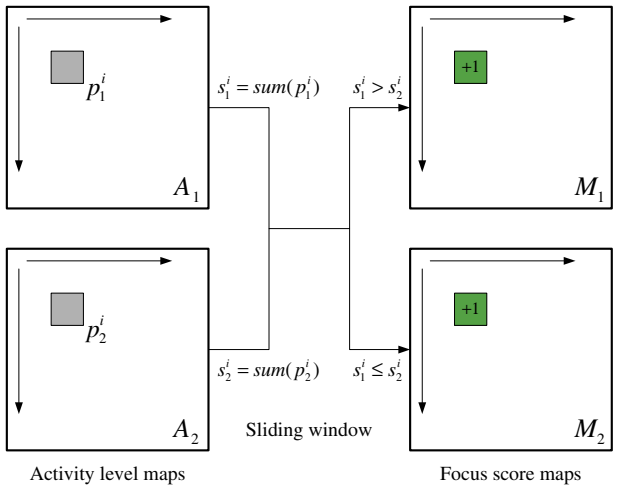


Fig. 4. The construction of focus score maps with the sliding window technique.

Furthermore, it can be found that the above two rules are complementary, which means that if $I_1(x,y)$ is a focused pixel, then $I_2(x,y)$ must be a defocused pixel, and vice versa. Moreover, the uncertain pixels in I_1 and I_2 have the same positions where either $M_1(x,y)$ or $M_2(x,y)$ equals to 0. Since the above classification criterion has a strong constraint to make a definite decision, the obtained focused/defocused results are usually reliable. Fig. 5(a) and (b) show the classification maps of the two source images in Fig. 3(a) and (d), respectively. For each initial classification map, the focused pixels are assigned to "1" (white) while other pixels (defocused and uncertain) are assigned to "0" (black).

As an initial classification map usually contains some small focused regions as well as some small holes surrounded by the focused region, we applied a simple post-processing approach to remove these regions. Specifically, a focused region which is smaller than a given area threshold is removed, while a hole which is smaller than the same threshold is filled as focused region. The area threshold is adaptively set to the one-hundredth of the total number of pixels in one source image. The post-processed segmentation maps of Fig. 5(a) and (b) are shown in Fig. 5(c) and (d), respectively. Fig. 5(e) shows the initial decision map T, which is obtained by

$$T(x,y) = \begin{cases} 1, & \text{if } I_1(x,y) \text{ is a focused pixel} \\ 0, & \text{if } I_2(x,y) \text{ is a focused pixel} \\ 0.5, & \text{otherwise} \end{cases}. \tag{4}$$

In Fig. 5(e), the white "1" and the black "0" pixels represent the definite focused regions of the first and second source images, respectively. The gray "0.5" pixels indicate the indefinite regions, which will be further processed in the next subsection.

3.2. Feature matching and map refinement

As shown in Fig. 5(e), the indefinite regions tend to appear at the boundaries between focused and defocused regions. Moreover, when there are moving objects in the scene, the motion regions are also likely to be detected as indefinite regions. In this subsection, we take advantage of the feature matching ability of SIFT descriptor to make a further classification for the uncertain pixels in the initial decision map. In addition, the spatial frequency (SF) [33] is employed as a local focus measure in the refinement process. The SF is defined as

$$SF = \sqrt{RF^2 + CF^2}, \tag{5}$$

where $RF = \sqrt{\frac{1}{N_1 \times N_2} \sum_{n_1=0}^{N_1-1} \sum_{n_2=1}^{N_2-1} [I(n_1,n_2) - I(n_1,n_2-1)]^2}$ is row frequency, $CF = \sqrt{\frac{1}{N_1 \times N_2} \sum_{n_1=1}^{N_1-1} \sum_{n_2=0}^{N_2-1} [I(n_1,n_2) - I(n_1-1,n_2))]^2}$ is column frequency, and I is the input image of size $N_1 \times N_2$. Generally, for an image patch, a larger SF value indicates a higher activity level.

As mentioned before, the contents of multiple source images at these uncertain positions are usually not exactly the same. As shown in Fig. 6(a), the (x,y) denotes an uncertain position. For the pixel at (x,y) in each source image, by applying the normalized dense SIFT descriptors, we can search for its most matching pixel within a local window centered at (x,y) in the other source image. The Euclidean distance between two descriptors in the 128D space is used to measure the similarity of two corresponding pixels, and the nearest neighbor is selected as the matching pixel. Since the source images have been pre-registered as well considering the computational efficiency, a 7×7 window is used to limit the search space in our method. Let (x_1,y_1) denotes the matching pixel of $I_1(x,y)$ in I_2 while (x_2,y_2) denotes the matching point of $I_2(x,y)$ in I_1, as shown in Fig. 6(a). To obtain the final decision map $D(x,y)$, only the indefinite regions in $T(x,y)$

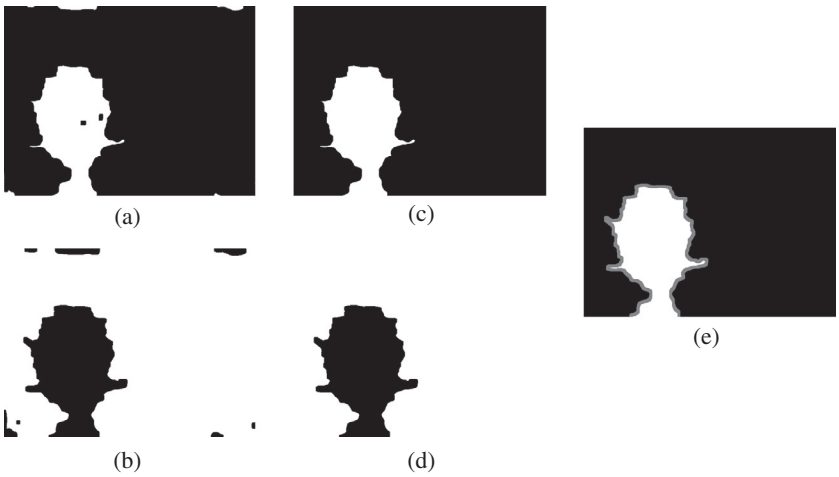


Fig. 5. The initial segmentation results. (a and b) Initial classification maps, (c and d) the post-processed segmentation maps, and (e) the initial decision map.

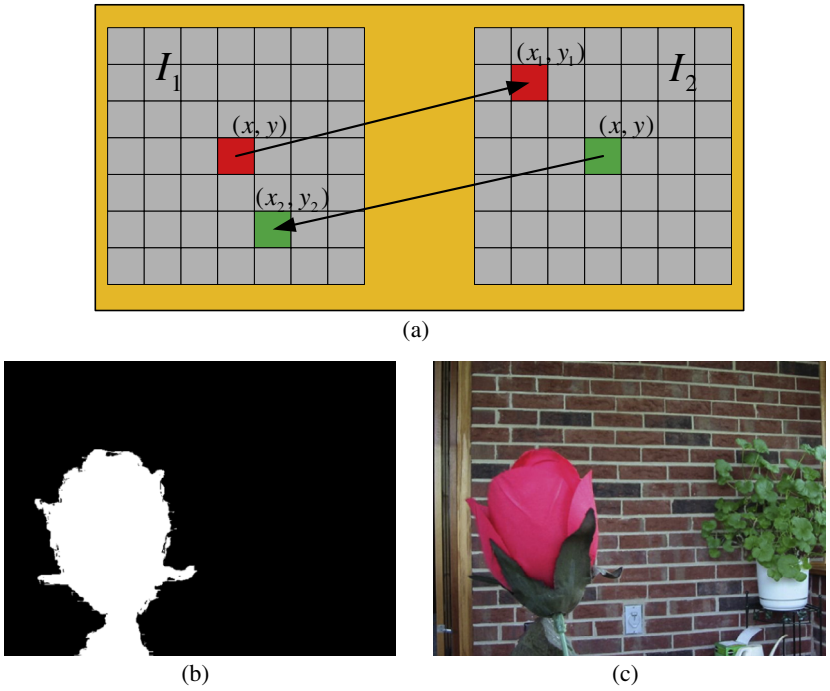


Fig. 6. Feature matching and map refinement. (a) Diagram of feature matching, (b) final decision map, and (c) the fused image obtained by the proposed algorithm.

(i.e., $T(x,y) = 0.5$) are further processed while the values of definite regions in $T(x,y)$ (i.e., $T(x,y) = 1$ or $T(x,y) = 0$) remain unchanged in $D(x,y)$. Particularly, via feature matching and spatial frequency, the pixel at (x,y) in the indefinite regions in $T(x,y)$ is refined with the following rule

$$D(x,y) = \begin{cases} 1, & \text{if } SF(w_1(x,y)) > SF(w_2(x_1,y_1)) \\ & \text{and } SF(w_2(x,y)) < SF(w_1(x_2,y_2)) \\ 0, & \text{if } SF(w_1(x,y)) < SF(w_2(x_1,y_1)) \\ & \text{and } SF(w_2(x,y)) > SF(w_1(x_2,y_2)) \\ 0.5, & \text{otherwise} \end{cases}, \tag{6}$$

where $w_1(x,y)$ and $w_2(x,y)$ denote a local window centered at (x,y) in I_1 and I_2, respectively. In our algorithm, the size of this window is also set to 7×7. This rule means that only when $I_1(x,y)$ is clearer than $I_2(x_1,y_1)$ while $I_2(x,y)$ is more blurry than $I_1(x_2,y_2)$, the pixel at (x,y) will be viewed as a focused pixel in I_1. Correspondingly, only when $I_2(x,y)$ is clear than $I_1(x_2,y_2)$ while $I_1(x,y)$ is more blurry than $I_2(x_1,y_1)$, the pixel at (x,y) will be viewed as a focused pixel in I_2. Otherwise, the focus information at (x,y) still remains uncertain and $D(x,y)$ is set to 0.5. Fig. 6(b) shows the obtained final decision map. We can see that most of the uncertain pixels in Fig. 5(e) are correctly classified.

3.3. Fusion

With the final decision map D, the fused image I_F is calculated by

$$I_F(x,y) = D(x,y)I_1(x,y) + (1 - D(x,y))I_2(x,y). \tag{7}$$

The fused image is shown in Fig. 6(c).

3.4. Generalized fusion algorithm for multiple source images

The algorithm presented above only aims at fusing two source images. However, it can be straightforwardly extended when there

are three or more source images. Both the classification rule in Eq. (2) and the refinement rule in Eq. (6) can be easily extended to process multiple source images. Specifically, the generalized algorithm for fusing N pre-registered source images $I_i, i \in \{1,\ldots,N\}$ is described as follows:

Step 1: For each source image, first calculate its dense SIFT image (unnormalized), then obtain its activity level map by accumulating all the elements in a SIFT descriptor, and finally normalize its original SIFT image.
Step 2: Via the sliding window technique, compare all the activity level maps with "choose-max" strategy to obtain N focus score maps $M_i, i \in \{1,\ldots,N\}$, as described in Section 3.1.
Step 3: For each source image, search for its definite focused pixels with the focus score maps. A pixel at (x,y) in I_i will be classified as a focused pixel if $\sum_{j=1,j\neq i}^{N} M_j(x,y) = 0$ is valid. Obviously, the classification rule in Eq. (2) is its special case when $N = 2$.
Step 4: Apply the post-processing technique introduced in Section 3.1 to obtain the definite focused regions for each source image. The uncertain pixels which do not belong to any of the focused regions construct the indefinite regions. Thus, an initial decision map is obtained for each source image.
Step 5: For every uncertain pixel in each source image, find its most similar pixel among the local neighborhoods in all the other source images, i.e., find only one correspondence in the $7 \times 7 \times (N-1)$ search space. Therefore, there are totally N matched pairs.
Step 6: Compare the local activity level of the N matched pairs using spatial frequency. For each source image, if the activity level of an uncertain pixel is higher than its correspondence, it will be classified as a focused pixel. Otherwise, it will be classified as a defocused pixel. Thus, there are three possible situations for a pixel at (x,y).

(i) If a pixel at (x,y) is classified as a focused pixel only in one source image, the final value at (x,y) in the decision map of this source image is set to 1, while the final value at (x,y) in each of the other decision maps is set to 0.
(ii) If a pixel at (x,y) is simultaneously classified as a focused pixel in more than one source images, the final value at (x,y) in each of the decision maps of these source images is averaged by the number of these images, while the final value at (x,y) in each of the other decision maps is set to zero.
(iii) If a pixel at (x,y) is classified as a defocused pixel in all the source images, the final value at (x,y) in every decision map is set to $1/N$.

Based on the above rule, for each source image, a final decision map $D_i, i \in \{1,\ldots,N\}$ is obtained. It is not difficult to find that when there are only two source images, this step is exactly the same as the refinement approach presented in Section 3.2.

Step 7: Obtain the fused image with $I_F(x,y) = \sum_{i=1}^{N} D_i(x,y) I_i(x,y)$.

4. Experiments

In this section, we first present the detailed information of experimental settings, and then analyze the influences of two main parameters on fusion performance. Finally, the experimental results are exhibited and discussed.

4.1. Experimental settings

4.1.1. Source images

In our experiments, 12 pairs of popular multi-focus source images shown in Fig. 7 are mainly utilized to verify the effectiveness of the proposed fusion algorithm. Among them, six pairs are gray images while the other six pairs are color images. For consistency, we make an arrangement that the first (left) one of each pair

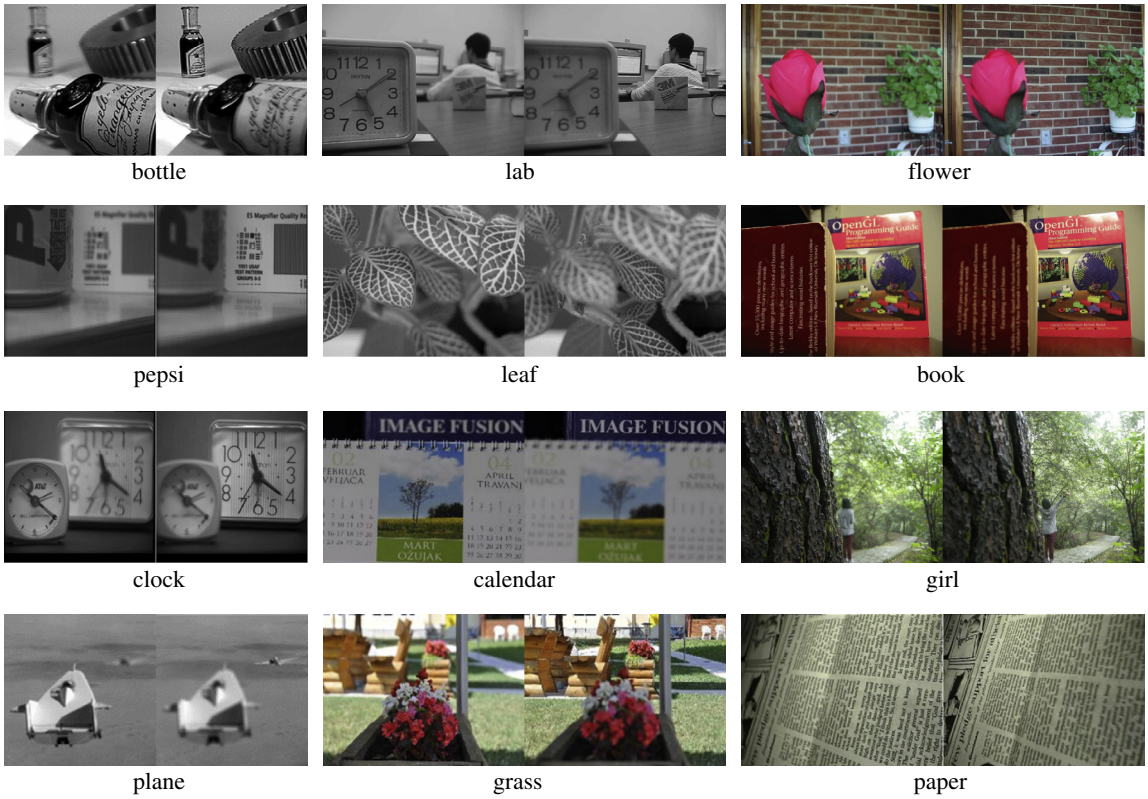


Fig. 7. Twelve pairs of multi-focus source images used in our experiments.

focuses on the near objects while the second (right) one focuses on far objects.

4.1.2. Compared methods

To confirm the effectiveness of the proposed dense SIFT-based image fusion methods (DSIFT), nine representative fusion methods which are based on discrete wavelet transform (DWT) [7], dual tree complex wavelet (DTCWT) [8], non-subsampled contourlet transform (NSCT) [10], neighbor distance (ND) [13], spatial frequency-motivated pulse coupled neural networks in non-subsampled contourlet transform domain (NSCT–PCNN) [11], sparse representation (SR) [15], higher order singular value decomposition (HOSVD) [16], image matting (IM) [23] and guided filtering (GF) [24], are selected to make comparisons. According to the classification introduced in Section 1, the DWT, DTCWT, NSCT and ND methods are multi-scale transform methods, the SR and HOSVD methods are feature space transform methods, the IM and GF methods are spatial methods, and the NSCT–PCNN method is a PCNN-based as well as a multi-scale transform method.

The parameters of the above nine methods are set as follows. For the DWT, DTCWT, NSCT and ND methods, the levels of decomposition are all set to 4, the low-frequency coefficients adopt the "average" scheme while the high-frequency coefficients adopt the "max-absolute" scheme. For the NSCT and NSCT–PCNN methods, the direction numbers of the four decomposition levels are selected as 4, 8, 8, 16, respectively. The other main parameters of NSCT–PCNN, ND, SR, HOSVD, IM and GF methods are set as the recommended values which are reported in the respective publications. In our experiments, the DWT, DTCWT and NSCT methods are implemented based on O. Rockinger's image fusion toolbox [34] and some multi-scale transform toolboxes downloaded from MATLAB Central [35]. The code of the NSCT–PCNN method is available on the X. Qu's homepage [36]. The codes of the IM and GF methods are available on X. Kang's homepage [37].

4.1.3. Objective evaluation metrics

Quantitative evaluation of the image fusion performance is not an easy task since it is practically impossible to obtain the ground truth (reference) image. Although many metrics for assessing fusion performance have been introduced, none of them is definitely better than others. Therefore, it is necessary to employ multiple metrics to evaluate the performance of a fusion method. Liu et al. [38] proposed a comprehensive survey of evaluation metrics for image fusion. In [38], 12 popular fusion metrics are categorized into four classes, namely, information theory-based metrics, image feature-based metrics, image structural similarity-based metrics, and human perception-based metrics. In our experiments, we choose six of them covering all the four categories to confirm the effectiveness of our fusion method. For convenience, the six metrics are briefly introduced as follows. Uniformly, let A and B denote two source images of size $H \times W$ while F the fused image.

(i) Normalized mutual information Q_{MI} [39]. Q_{MI} is a fusion metric based on information theory. Hossny et al. proposed Q_{MI} to overcome the instability of traditional mutual information (MI)[40]. Q_{MI} is defined as

$$Q_{MI} = 2\left[\frac{MI(A,F)}{H(A)+H(F)} + \frac{MI(B,F)}{H(B)+H(F)}\right], \tag{8}$$

where $H(X)$ is the entropy of image X and $MI(X,Y)$ is the mutual information between image X and Y. The detailed definitions can be found in [39]. The metric Q_{MI} measures the amount of information in the fused image inherited from the source images.

(ii) Nonlinear correlation information entropy Q_{NCIE} [41]. The Q_{NCIE} proposed by Wang et al. is also an information theory-based metric, which appears in the Chapter 19 in the book [3] as well. Q_{NCIE} is defined as

$$Q_{NCIE} = 1 + \sum_{i=1}^{3} \frac{\lambda_i}{3} \log_{256} \frac{\lambda_i}{3}, \tag{9}$$

where $\lambda_i, i \in \{1,2,3\}$ are the eigenvalues of the nonlinear correlation matrix

$$R = \begin{pmatrix} 1 & NCC_{AB} & NCC_{AF} \\ NCC_{BA} & 1 & NCC_{BF} \\ NCC_{FA} & NCC_{FB} & 1 \end{pmatrix}, \tag{10}$$

where NCC_{XY} is the nonlinear correlation coefficient between image X and Y [41].

(iii) Gradient-based fusion metric Q_G [42]. Q_G is a commonly used fusion metric which evaluates the extent of gradient information injected into the fused image from the source images. Q_G is calculated by

$$Q_G = \frac{\sum_{x=1}^{H}\sum_{y=1}^{W}(Q^{AF}(x,y)w^A(x,y) + Q^{BF}(x,y)w^B(x,y))}{\sum_{x=1}^{H}\sum_{y=1}^{W}(w^A(x,y) + w^B(x,y))}, \tag{11}$$

where $Q^{AF}(x,y) = Q_g^{AF}(x,y)Q_\alpha^{AF}(x,y)$, while $Q_g^{AF}(x,y)$ and $Q_\alpha^{AF}(x,y)$ denote the edge strength and orientation preservation values at pixel (x,y). The definition of $Q^{BF}(x,y)$ is similar. The weighting factors $w^A(x,y)$ and $w^B(x,y)$ indicate the significance of $Q^{AF}(x,y)$ and $Q^{BF}(x,y)$, respectively.

(iv) Phase congruency-based fusion metric Q_P [43]. Q_P is also a feature-based metric proposed by Zhao et al. The principal moments of the image phase congruency are used to define Q_P since they contain the information of image salient features such as edges and corners. The definition of Q_P is

$$Q_P = (P_p)^\alpha (P_M)^\beta (P_m)^\gamma, \tag{12}$$

where p, M and m refers to phase congruency, maximum and minimum moments, respectively. The detailed definition of Q_P can be found in [43]. The three exponential parameters α, β and γ are all set to 1 in our experiments.

(v) Yang's fusion metric Q_Y [44]. Q_Y is a structural similarity-based fusion metric, which measures the level of structural information of source images preserved in the fused image. The definition of Q_Y is

$$Q_Y = \begin{cases} \lambda(w)SSIM(A,F\,|\,w) + (1-\lambda(w))SSIM(B,F\,|\,w), & SSIM(A,B\,|\,w) > 0.75 \\ \max\{SSIM(A,F\,|\,w), SSIM(B,F\,|\,w)\}, & SSIM(A,B\,|\,w) < 0.75 \end{cases}, \tag{13}$$

where $SSIM$ is the structural similarity [45], w is a 7×7 window and $\lambda(w)$ is the local weight [46] calculated by

$$\lambda(w) = \frac{s(A\,|\,w)}{s(A\,|\,w) + s(B\,|\,w)}, \tag{14}$$

where $s(X\,|\,w)$ is the variance of image X within the window w.

(vi) Chen-Blum metric Q_{CB} [47]. Q_{CB} is a human perception-based fusion metric which utilizes the major features in the human visual system model. The calculation of Q_{CB} is complex and more details can be found in [47].

In order to guarantee the fairness and objectivity of evaluation results, all the above six fusion metrics are implemented by a third-party: the image fusion evaluation toolbox provided by Professor Z. Liu who is the first author of [38]. Furthermore, the default parameters reported in the related publications are used

in our experiments. For all the six evaluation metrics, a larger value indicates a better fused result.

4.2. Analysis of algorithm parameters

The proposed fusion algorithm has two free parameters which are (1) the scale factor (the size of neighborhood patch) in dense SIFT calculation and (2) the block size of sliding window used in activity level comparison. In this subsection, we apply the above six objective metrics to evaluate the impacts of these two parameters on fusion performance. For each metric, the average value of all the twelve pairs of source images shown in Fig. 7 is employed for evaluation. First, we fix the block size to 8×8 when investigating the influence of scale factor. Then, when evaluating the impact of block size, the scale factor is fixed to 48.

The analysis result is shown in Fig. 8. To have a better observation of both global and local changing trends, for each of (a) and (b) in Fig. 8, we not only give the left graph in which all the six metrics are displayed in a larger scale vertical axis, but also provide six curves separately in smaller scale axes on the right side. From Fig. 8(a), we can see that the values of all the metrics improve as the scale factor increase before it reaches 48. When the scale factor is larger than 48, some metrics such as Q_G and Q_P decrease as the scale factor increase, and some metrics such as Q_{MI}, Q_{NCIE}, Q_Y and Q_{CB} tend to be stable. Considering that the computational efficiency of dense SIFT will decrease when the scale factor becomes larger, it is a reasonable choice to set the scale factor to 48. On the other hand, it can be seen from Fig. 8(b) that although the Q_{MI} and Q_{NCIE} decrease as the block size increase, the values of Q_G, Q_P, Q_Y and Q_{CB} are all the highest ones when the block size is 8×8. As the information theory-based metrics are not always very objective to evaluate the fusion performance alone [24], it is reasonable to set the block size to 8×8. Furthermore, from the left two integrated graphs in Fig. 8, we can find that the performance of the DSIFT fusion method is not sensitive to parameter choice. Therefore, the scale factor is fixed to 48 while the block size is fixed to 8×8 for all the test images in our experiments.

4.3. Experimental results and discussions

4.3.1. Comparisons with other fusion methods

The "flower", "girl" and "lab" source images and their fused images with different fusion methods are shown in Figs. 9–11, respectively. In the "flower" images, the edge of the flower is irregular and the chrominance of the brick wall in the two source images are not the same. In the "girl" images, there is a clear

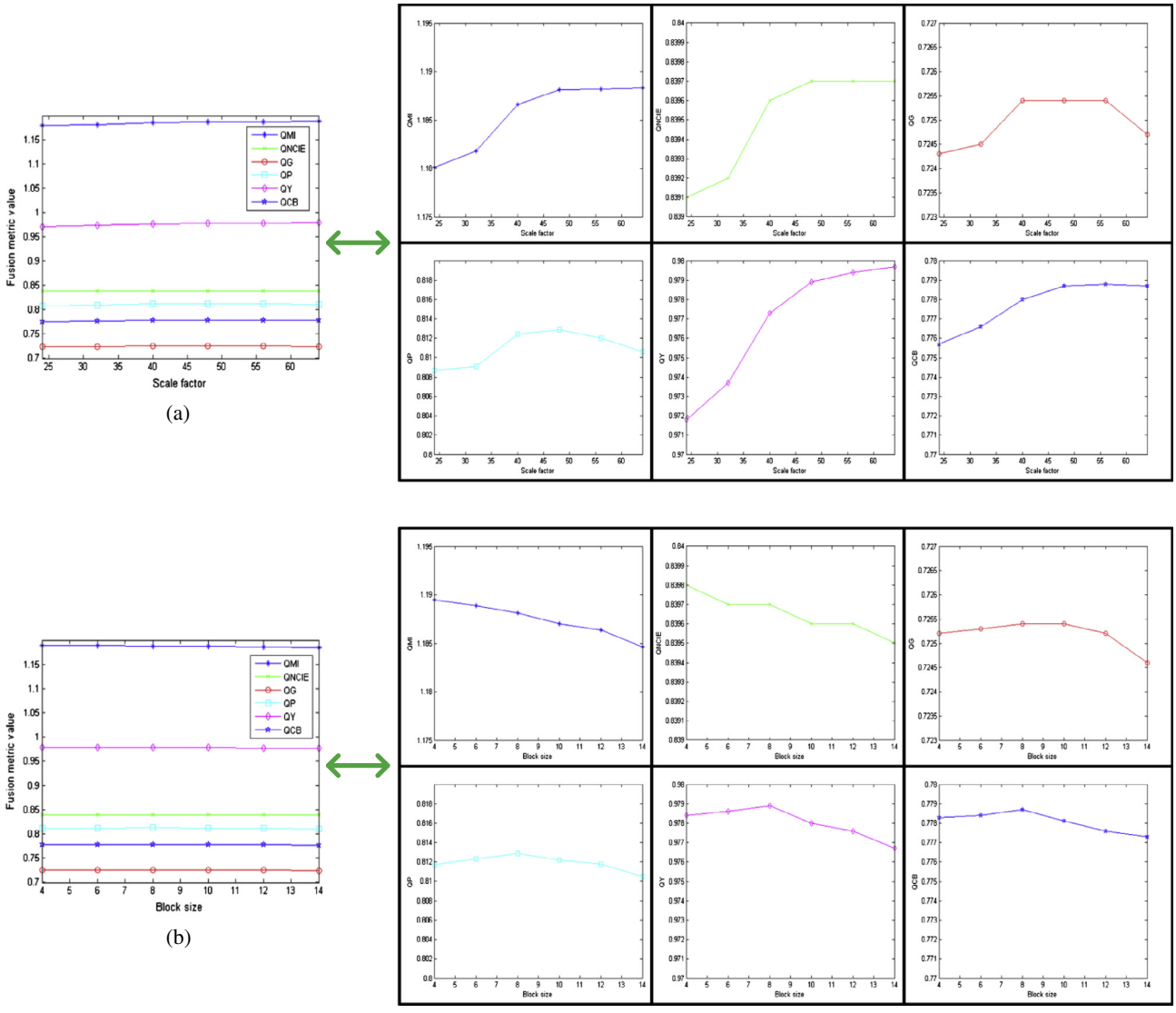
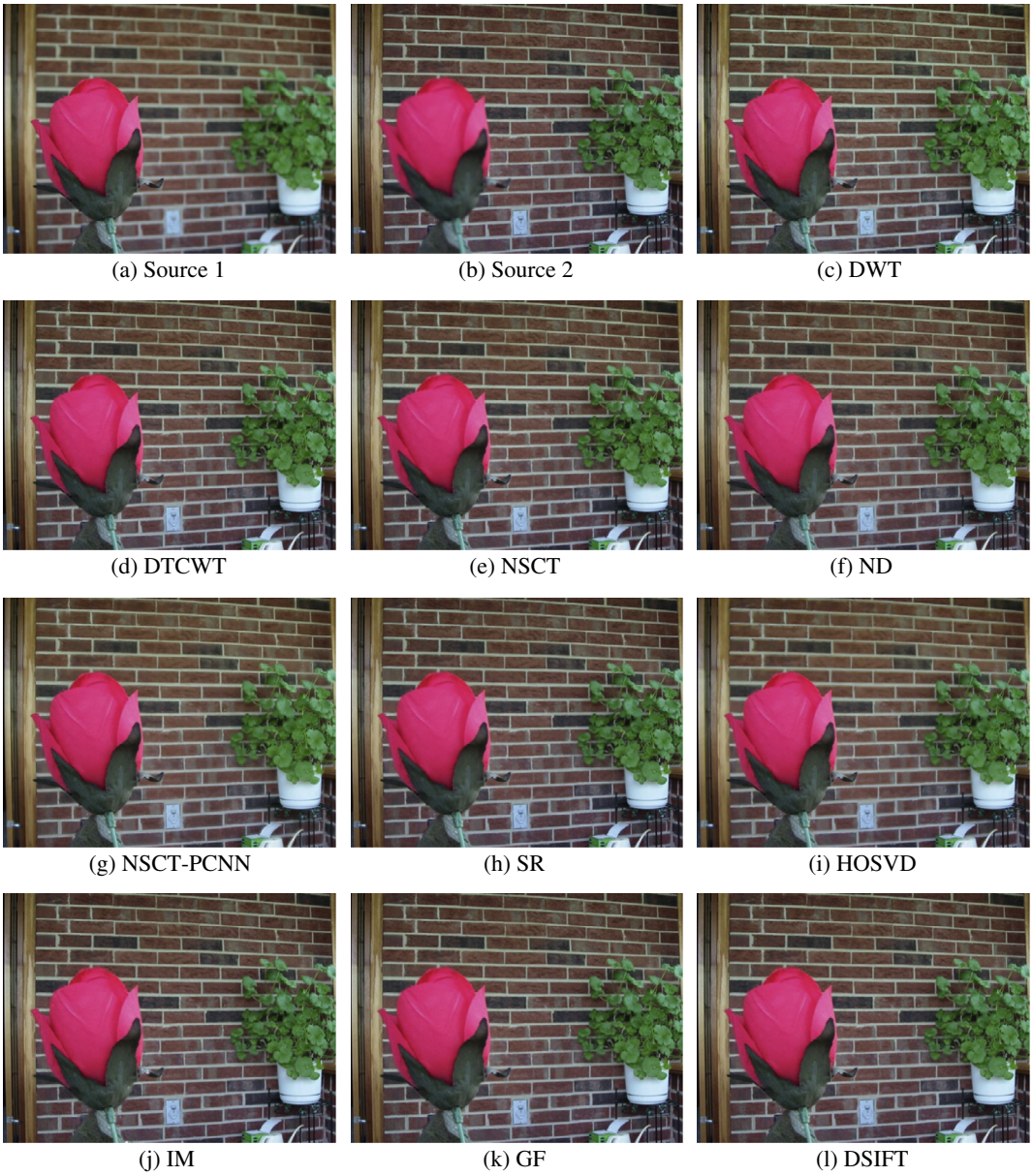


Fig. 8. Objective performance of the proposed method with different parameters. (a) Scale factor and (b) block size.

movement of the girl between two source images. In the classic "lab" images, the two source images also cannot be well registered as there is a slight motion of the student's head. To make a close-up view for better comparison, two magnified regions containing the moving objects are extracted from the "girl" and "lab" images, respectively, as shown in Figs. 12 and 13.

In each of the figures from Figs. 9–13, the subfigures (a) and (b) are two source images, and the fused results obtained by DWT, DTCWT, NSCT, ND, NSCT–PCNN, SR, HOSVD, IM, GF and DSIFT fusion methods are shown in subfigures (c)–(l), respectively. Generally, the fused results of DWT method suffer from serious artifacts, especially for the regions with moving objects since DWT is not shift-invariant (see Figs. 12(c) and 13(c)). The fused results of DTCWT and NSCT methods are better than those of the DWT method, but the artifacts in the motion regions are still obvious. The ND method can achieve better results than the NSCT method in the motion regions (see the girl's coat in Fig. 12(e) and (f)). The NSCT–PCNN method can achieve better results than the NSCT method in well registered regions, but the visual quality is even worse in the motion regions (see the student's head in Fig. 12(e) and (g)). The SR and HOSVD methods can usually achieve better results than the multi-scale transform methods, but the artifacts may still appear in the motion regions especially for the HOSVD method (see the girl's head in Fig. 12(i) and the student's head in Fig. 13(h) and (i)). The IM and GF methods can effectively prevent the artifacts from appearing in the motion regions. However, with careful observation, we can find that there are still some small defects of these two methods. The IM method tends to over-sharpen the boundaries between focused and defocused regions (see the edge of flower especially the receptacle in Fig. 9(j), the edge of trunk especially the part near the girl's left leg in Fig. 12(j), and the upper right corner of the clock in Fig. 11(j)). On the contrary, the GF method inclines to under-sharpen those boundaries, which means that the boundaries are likely to be little blurred (see the edge of flower especially the receptacle in Fig. 9(k), the edge of trunk in Fig. 12(k), and the student's head in Fig. 13(k)). Like the IM and GF methods, the proposed DSIFT fusion method can also obtain fused images without producing any artifacts and



(a) Source 1 (b) Source 2 (c) DWT

(d) DTCWT (e) NSCT (f) ND

(g) NSCT-PCNN (h) SR (i) HOSVD

(j) IM (k) GF (l) DSIFT

Fig. 9. The "flower" source images and fused images with different methods.

Fig. 10. The "girl" source images and fused images with different methods.

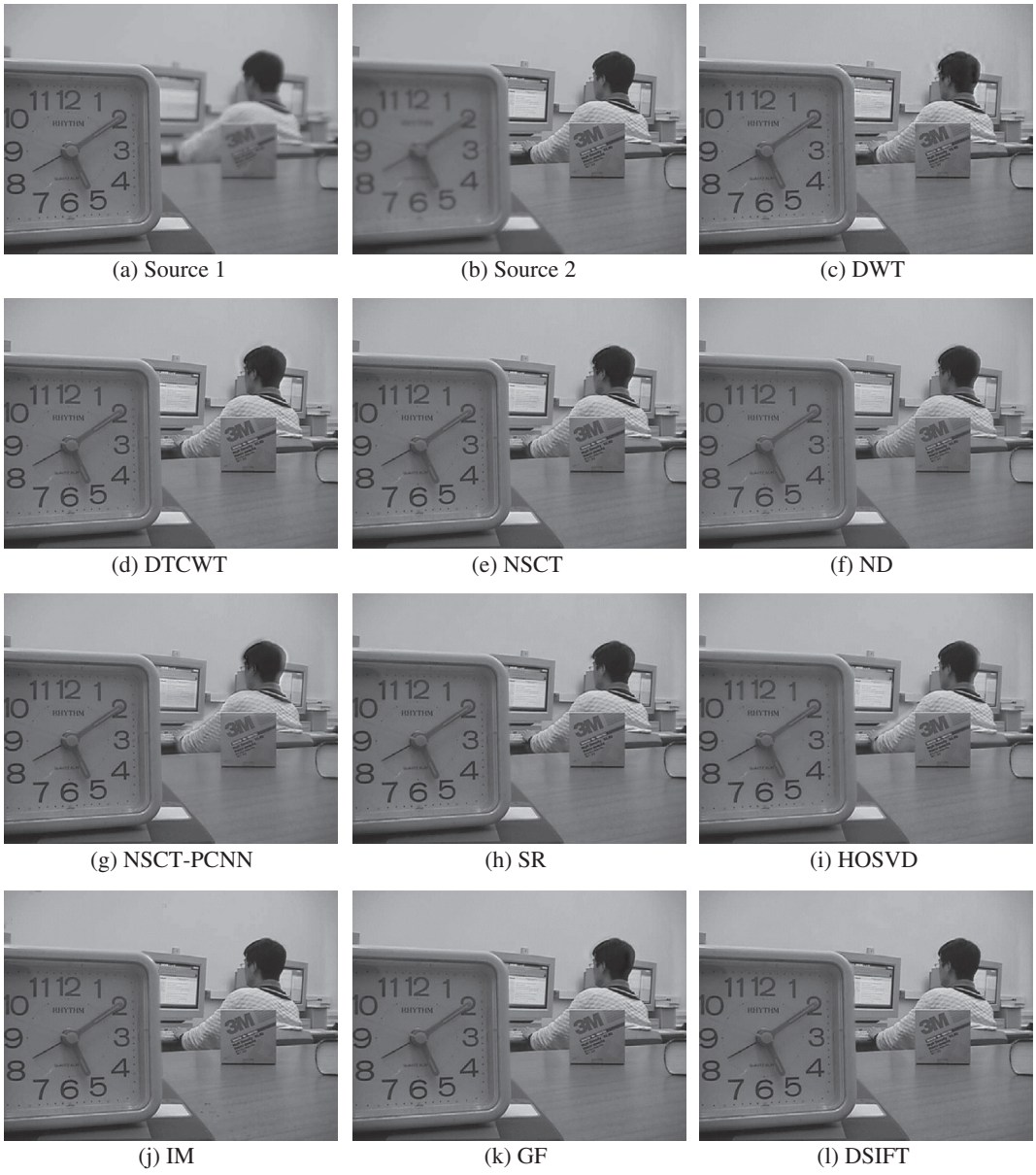
color distortions. Furthermore, it can process the boundary regions better than either IM or GF method and the fused boundaries are more natural (see the edge of flower especially the receptacle in Fig. 9(l), the edge of trunk in Fig. 12(l), and the student's head in Fig. 13(l)). Fig. 14 shows the fused results of other nine sets of images, and only the fused images of IM, GF and DSIFT methods are presented because of space limitation. We can see that the proposed method can competitive with the other two methods which can generally represent the state of the art.

In addition to the visual comparisons, the six objective evaluation metrics, i.e., $Q_{MI}, Q_{NCIE}, Q_G, Q_P, Q_Y$, and Q_{CB} are employed to quantitatively compare the performance of different fusion methods. The average assessment values of twelve sets of test images are listed in Table 1, in which the highest value is shown in bold and the numbers in parentheses denote the number that the related method gets the first place. From Table 1, we can clearly see that the proposed DSIFT beat all the nine compared methods in terms of all the six metrics. Although a single metric may not always reflect the fused quality objectively, the fact that the advantages on all the six metrics which cover the entire four categories presented in [38] can definitely confirm the effectiveness of the proposed method. Furthermore, for each of the six metrics in Table 1, the number in the parenthesis of our method is larger than those of other methods, which can reflect the stability and reliability of the proposed fusion method to some extent.

4.3.2. Matching vs. non-matching

To verify the effectiveness of feature matching step in our fusion algorithm, we apply a new dense SIFT-based method without feature matching for comparison. The only difference between the new compared method and the original DSIFT method is that the new version directly compares the local spatial frequency of two uncertain pixels owning the same position, i.e., the process of feature matching is not carried out. Thus, the refinement rule for the pixel at (x, y) in the indefinite regions in Eq. (6) is replaced by

$$D(x,y) = \begin{cases} 1, & \text{if } SF(w_1(x,y)) > SF(w_2(x,y)) \\ 0, & \text{if } SF(w_2(x,y)) > SF(w_1(x,y)), \\ 0.5, & \text{otherwise} \end{cases} \tag{15}$$



(a) Source 1 (b) Source 2 (c) DWT

(d) DTCWT (e) NSCT (f) ND

(g) NSCT-PCNN (h) SR (i) HOSVD

(j) IM (k) GF (l) DSIFT

Fig. 11. The "lab" source images and fused images with different methods.

where $w_1(x,y)$ and $w_2(x,y)$ have the same meanings as defined in Eq. (6). In addition, all the other parameters in the compared method remain unchanged. Fig. 15 shows an example of the fused result with/without feature matching. We can see that the artifacts around the student's head in the non-matching fused image are relatively serious (see Fig. 15(a) and (c)). This is mainly because there exists a slight motion of the student's head between the two source images. However, with feature matching, these artifacts can be successfully eliminated (see Fig. 15(b) and (d)).

4.3.3. Comparison of computational efficiency

The computational efficiency of different fusion methods is compared here. In our experiments, all the ten test methods are implemented in MATLAB on a computer with a 3.0 GHz CPU and 4 GB RAM. For two source images of size 640×480, the average running time of different fusion methods is listed in Table 2. It can be seen that the SR method is the most time consuming one while the HOSVD and NSCT–PCNN methods are also not very efficient. The IM and GF methods both have a high computational efficiency, especially the latter one. The DSIFT method is in the middle level and requires about 15 s to fuse two source images of size 640×480. We believe that with a more efficient implementation approach such as C++, the running time can be easily reduced to less than one second, which can meet the requirement of many applications. Furthermore, experiments shows that about 53% of the total running time is consumed at the feature matching step. When we apply the method in the non-matching mode described in Section 4.3.2, the average running time for two 640×480 source images is about 7 s. After all, this fusion mode can still obtain fused results with high quality if the two source images are well pre-registered.

4.3.4. More fusion examples

Fig. 16 shows a fusion example of three oil painting source images downloaded from website [48]. The three source images shown in Fig. 16(a)–(c) are near focused, middle focused and far focused, respectively. The fused results of IM, GF and the proposed DSIFT methods are shown in Fig. 16(d)–(f), respectively. For three

Fig. 12. Magnified regions of the "girl" source images and fused images with different methods.

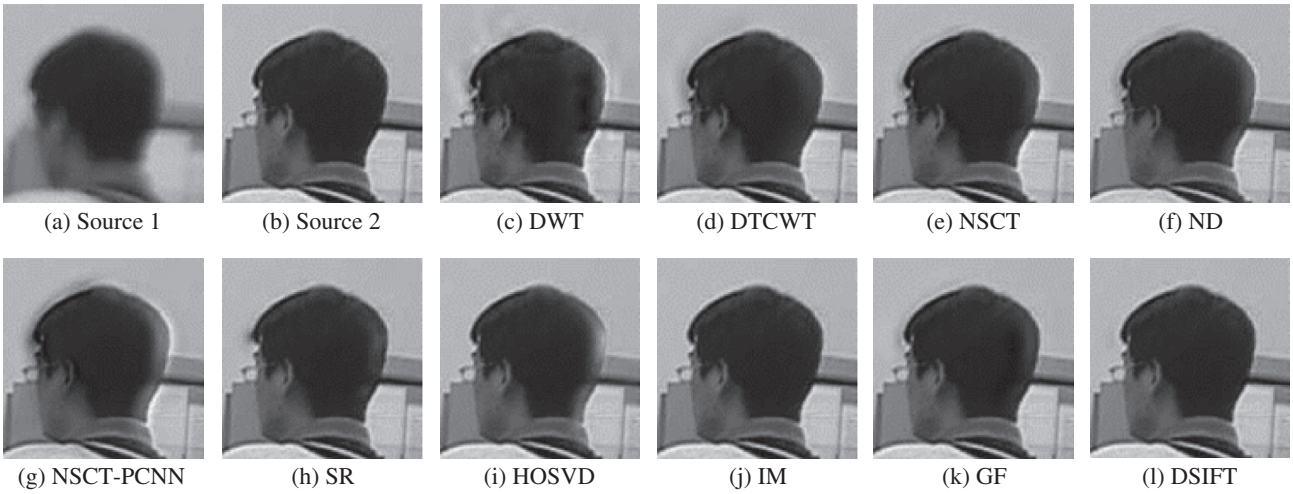


Fig. 13. Magnified regions of the "lab" source images and fused images with different methods.

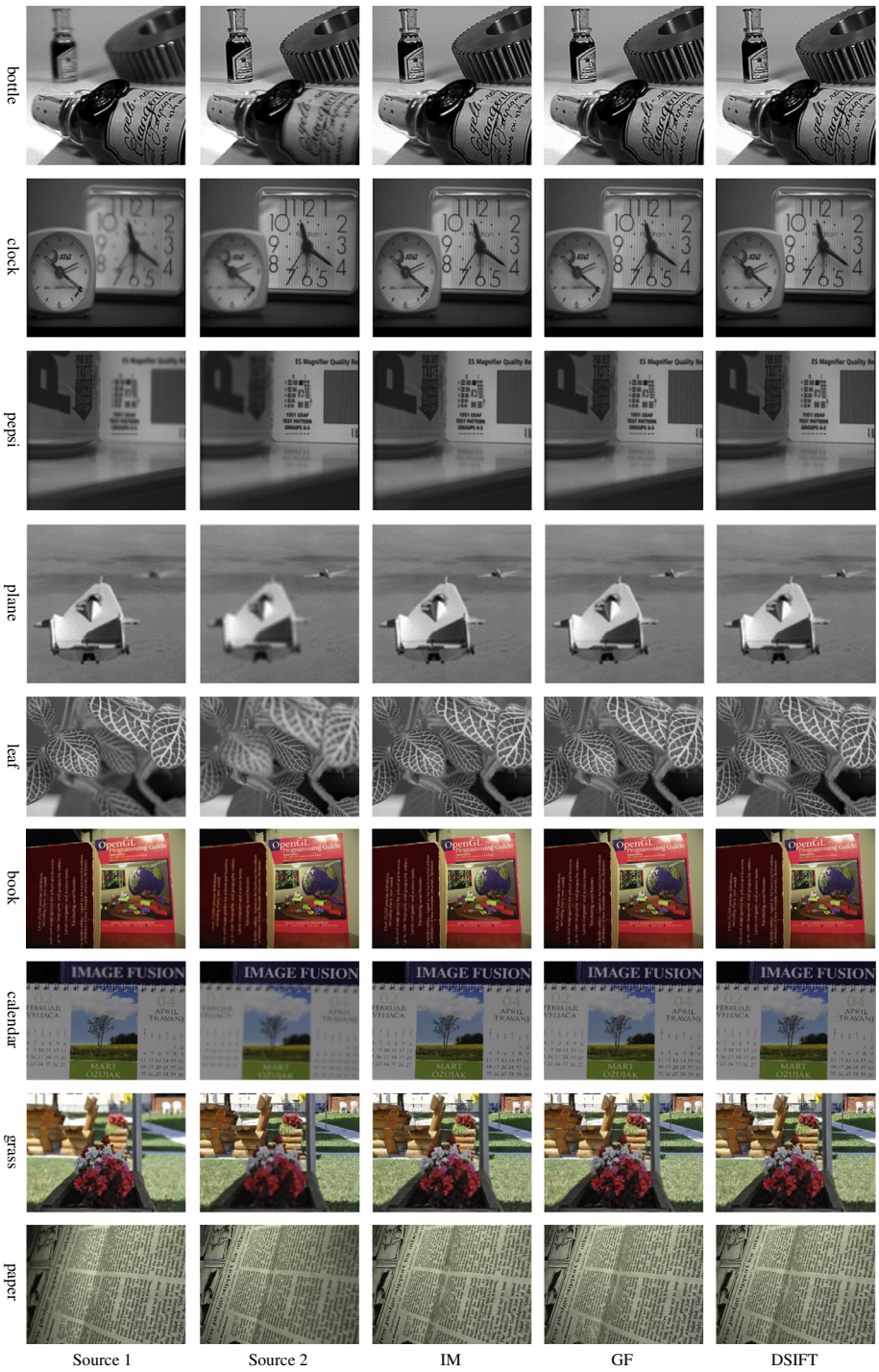


Fig. 14. Other source images and fused results obtained by IM, GF and DSIFT methods.

source images, unlike the IM or GF method fuses them in series, i.e., the first and second images are merged at first, and then the third image is merged with the above result to obtain the final fused image, we use the generalized DSIFT fusion algorithm presented in Section 3.4 to merge them in parallel. We can see from Fig. 16(f) that all the useful information of three source images are transferred to our fused image.

Fig. 17 exhibits a fusion example of a multi-focus image sequence downloaded from website [49]. As shown in Fig. 17(a), there are 20 "beijing map" source images of size 720×480, and each of them focuses on a narrow stripe. Fig. 17(b) shows the reference image which is also included in the data set. The fused results of IM, GF and DSIFT methods are shown in Fig. 17(c)–(e), respectively. It should be noticed that the "non-matching" model proposed in Section 4.3.2 is applied for this example mainly based on the consideration of computational efficiency. When the number of source images becomes larger, the computational cost of feature matching step increases a lot but the cost of other steps only increases a little as the parallel mechanism is used in the generalized fusion algorithm presented in Section 3.4. Moreover, the 20 source images in this example are well pre-registered, so it is a better choice to use the "non-matching" model here. From Fig. 17, it

Table 1
The average objective assessments of different fusion methods.

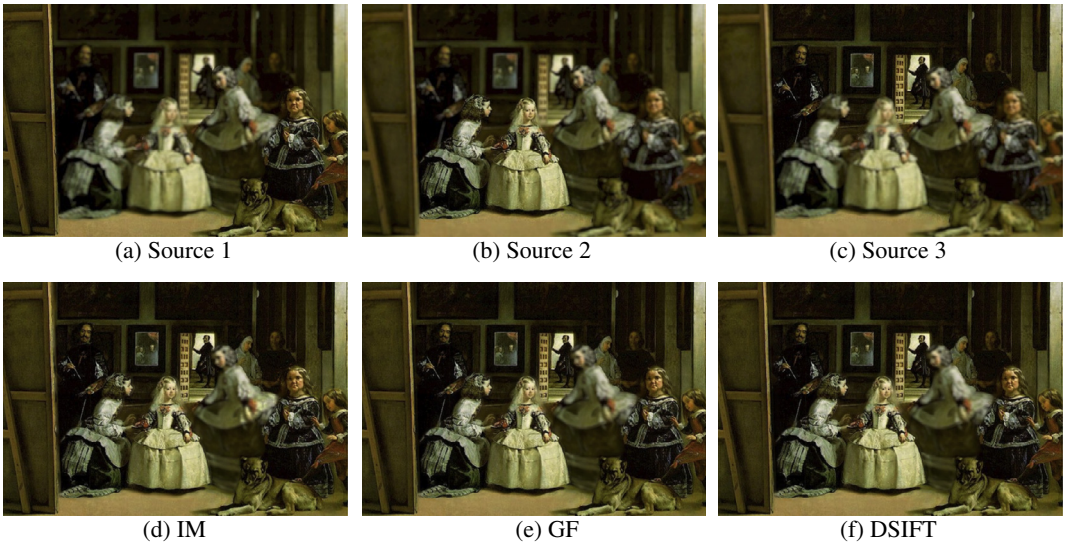
Method	DWT	DTCWT	NSCT	ND	NSCT–PCNN	SR	HOSVD	IM	GF	DSIFT
Q_{MI}	0.7473(0)	0.8336(0)	0.8530(0)	0.8532(0)	0.9811(0)	1.0512(0)	1.0314(0)	1.1559(2)	1.0278(0)	**1.1881(10)**
Q_{NCIE}	0.8193(0)	0.8224(0)	0.8232(0)	0.8235(0)	0.8312(0)	0.8322(1)	0.8305(0)	0.8383(2)	0.8315(0)	**0.8397(9)**
Q_G	0.6301(0)	0.6889(0)	0.6930(0)	0.6959(0)	0.6944(0)	0.7105(0)	0.7047(0)	0.7222(3)	0.7194(1)	**0.7254(8)**
Q_P	0.6702(0)	0.7626(0)	0.7647(0)	0.7621(0)	0.7596(0)	0.7855(0)	0.7786(1)	0.8001(0)	0.8068(3)	**0.8129(8)**
Q_Y	0.8699(0)	0.9229(0)	0.9325(0)	0.9378(0)	0.9415(0)	0.9480(0)	0.9539(1)	0.9770(5)	0.9627(0)	**0.9789(6)**
Q_{CB}	0.6777(0)	0.7187(0)	0.7295(0)	0.7039(0)	0.7307(0)	0.7398(0)	0.7330(0)	0.7716(1)	0.7568(0)	**0.7787(11)**

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

Fig. 15. An example of the fused result with/without feature matching. (a) The fused image with feature matching, (b) the fused image without feature matching, (c) the magnified regions extracted from (a), and (d) the magnified regions extracted from (b).

Table 2
Average running time of different methods on two source images of size 640 × 480.

Method	DWT	DTCWT	NSCT	ND	NSCT–PCNN	SR	HOSVD	IM	GF	DSIFT
Time/s	0.39	1.09	12.0	8.27	186	1637	99.2	5.92	1.12	15.4



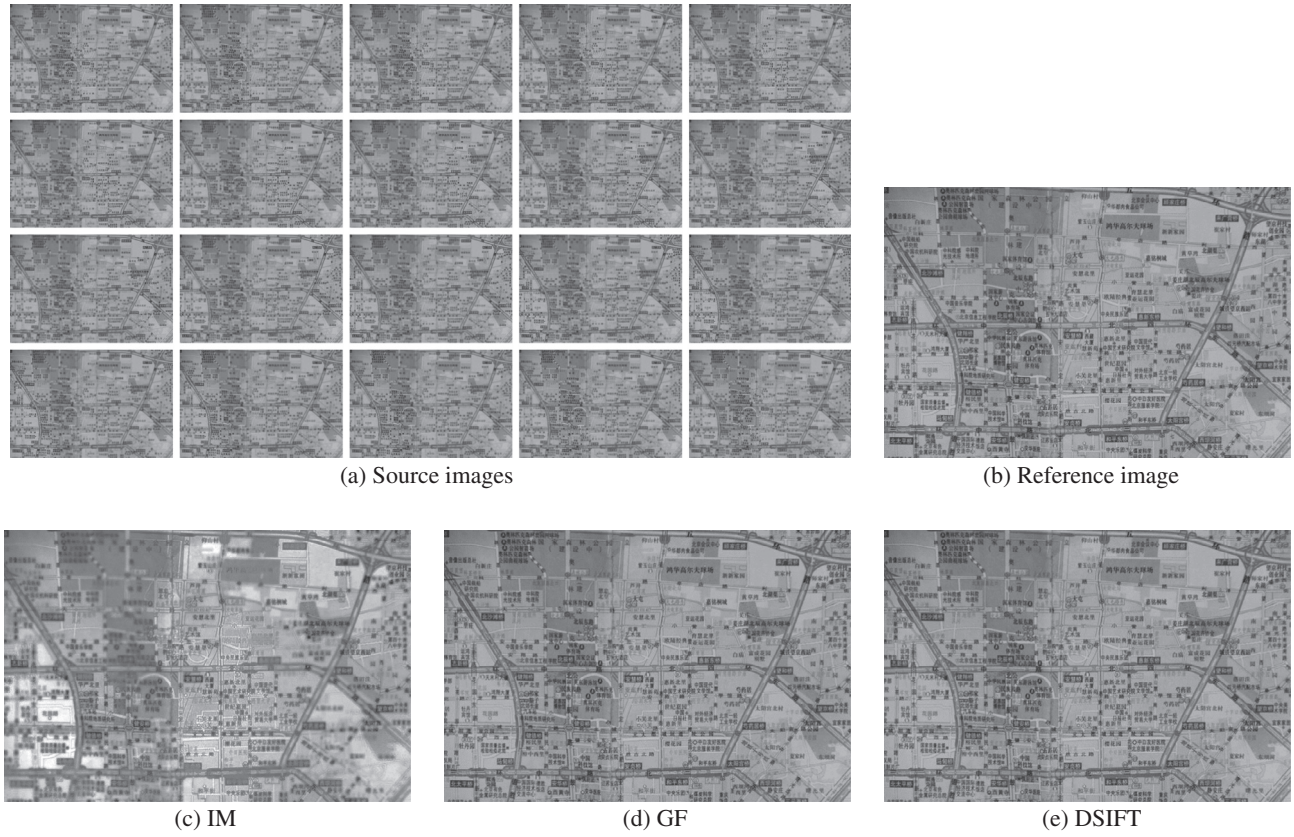
(a) Source 1 (b) Source 2 (c) Source 3

(d) IM (e) GF (f) DSIFT

Fig. 16. A fusion example of three multi-focus source images.

can be seen that the fused images of the GF and DSIFT methods are both very close to the reference image, but the IM method loses effectiveness at this time. Furthermore, the root mean square error (RMSE) [50] and the SSIM [45] mentioned before are used to quantitatively evaluate the difference between each fused image and the reference image. A smaller RMSE measure indicates a better result, while the closer the SSIM value to 1, the better the fused image. Table 3 lists the quantitative assessments as well as the running time of these three methods. We can see that for both RMSE and SSIM, the performance of GF method is the best, but the difference between it with our method is not obvious. Furthermore, when fusing a large number of images, the computational efficiency of our method is relatively improved compared with both IM and GF methods, which can be verified by comparing Table 2 with Table 3. This is because both IM and GF methods merge the source images one by one while our method fuses all the images at the same time.

At last, Fig. 18 gives an example for which our method does not perform very well. The two source images as shown in Fig. 18(a) and (b) are not well registered. It can be seen that all the buildings



(a) Source images (b) Reference image

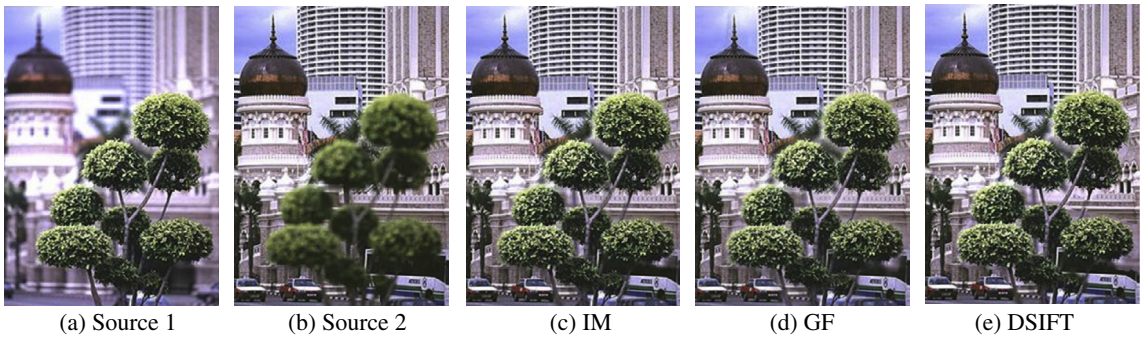
(c) IM (d) GF (e) DSIFT

Fig. 17. A fusion example of a multi-focus image sequence.

Table 3
the quantitative assessments and running time of three methods on the image sequence in Fig. 17.

Method	IM	GF	DSIFT
RMSE	21.7846	5.2029	5.3221
SSIM	0.7477	0.9547	0.9525
Time/s	130	24.7	56.6

in the second source image "move" a little to the right referring to the first source image, which may be caused by camera movement. Moreover, some vehicles in the scene were moving during the two captures. The fused results of IM, GF and the proposed methods are shown in Fig. 18(c)–(e), respectively. The fused result of the IM method owns the highest quality here, while either the GF method or the proposed method does not perform well enough at this time. For the fused image of the GF method, artifacts tend to appear around the edges of some objects, such as the spire of the upper left building in the scene. For our fused result, the main defects are on the small holes inside the plants in the foreground. Since the "hole-filling" strategy is used in our post-processing technique in Section 3.1, these small holes will be filled and falsely segmented as the foreground here. Thus, some details are missed in our fused image. In this situation, a post-processing approach with some "mild" smoothing strategies may be better, such as image filtering. However, these cases are minority after all. In most cases like the other examples given in this paper, the "hole-filling" strategy can work well and may more effective than smoothing strategies for the following two reasons. First, the smoothing strategies may be not powerful enough to correct all the mis-classified pixels or regions especially when a mis-classified region is not so small. Second, the object edges may be over-smoothed by a



(a) Source 1 (b) Source 2 (c) IM (d) GF (e) DSIFT

Fig. 18. A fusion example for which the proposed method does not perform well.

filtering-based strategy, which may lead some undesirable artifacts into the final fused images such as some results of the GF method mentioned in Section 4.3.1.

5. Conclusion and future work

In this paper, we propose a new multi-focus image fusion method with dense SIFT. In the proposed algorithm, via the sliding window technique, the dense SIFT descriptor is first used to measure the activity level of source image patches to obtain an initial decision map, and then the decision map is refined with SIFT feature matching and local focus measure comparison. The fusion method is also generalized to be capable of fusing more than two images. The proposed method is compared with nine representative fusion methods in terms of both visual perception and objective metrics. Experimental results demonstrate that the proposed fusion method can be competitive with or even outperform some state-of-the-art methods.

The main contribution of this paper is that it exhibits the great potential of image local features such as the dense SIFT used for image fusion. However, this work only makes a preliminary attempt and the proposed method also has some defects. For example, the computational efficiency is lower when compared with the guided filtering-based fusion method. Furthermore, the memory requirement is high for our method as the 128D feature descriptors need to be stored. Further research may be carried out in three aspects. The first is to construct more effective activity level measurements via image local feature descriptor. The second is to develop appropriate approaches such as dimension–reduction to make the fusion method less time or memory consuming. Finally, it is worthwhile to further investigate the practicability of the proposed method for other image fusion applications, such as multi-exposure image fusion and remote sensing image fusion.

Acknowledgements

The authors would first like to thank the editors and anonymous reviewers for their constructive and valuable comments and suggestions.

The authors would like to thank Dr. Miao Sun from University of Missouri-Columbia (USA) for his insightful suggestions. The authors would also like to express their thanks to Prof. Shutao Li and Dr. Xudong Kang from Hunan University (China), Prof. Zheng Liu from Toyota Technological Institute (Japan) for providing some source images and codes used in their publications [23,24,38]. This work is supported by the National Science and Technology Projects (No. 2012GB102007) and the National Natural Science Foundation of China (No. 61303150).

References

[1] S. Li, J. Kwok, I. Tsang, Y. Wang, Fusing images with different focuses using support vector machines, IEEE Trans. Neural Networks 15 (6) (2004) 1555–1561.
[2] A. Goshtasby, S. Nikolov, Image fusion: advances in the state of the art, Inf. Fus. 8 (2) (2007) 114–118.
[3] T. Stathaki, Image Fusion: Algorithms and Applications, Academic Press, 2008.
[4] G. Piella, A general framework for multiresolution image fusion: from pixels to regions, Image Fus. 4 (4) (2003) 259–280.
[5] P. Burt, E. Adelson, The Laplacian pyramid as a compact image code, IEEE Trans. Commun. 31 (4) (1983) 532–540.
[6] V. Petrovic, C. Xydeas, Gradient-based multiresolution image fusion, IEEE Trans. Image Process. 13 (2) (2004) 228–237.
[7] H. Li, B. Manjunath, S. Mitra, Multisensor image fusion using the wavelet transform, Graph. Models Image Process. 57 (3) (1995) 235–245.
[8] J. Lewis, R. OCallaghan, S. Nikolov, D. Bull, N. Canagarajah, Pixel- and region-based image fusion with complex wavelets, Inf. Fus. 8 (2) (2007) 119–130.
[9] F. Nencini, A. Garzelli, S. Baronti, L. Alparone, Remote sensing image fusion using the curvelet transform, Inf. Fus. 8 (2) (2007) 143–156.
[10] Q. Zhang, B. Guo, Multifocus image fusion using the nonsubsampled contourlet transform, Signal Process. 89 (7) (2009) 1334–1346.
[11] X. Qu, J. Yan, H. Xiao, Z. Zhu, Image fusion algorithm based on spatial frequency-motivated pulse coupled neural networks in nonsubsampled contourlet transform domain, Acta Automat. Sin. 34 (12) (2008) 1508–1514.
[12] S. Li, B. Yang, J. Hu, Performance comparison of different multi-resolution transforms for image fusion, Inf. Fus. 12 (2) (2011) 74–84.
[13] H. Zhao, Z. Shang, Y. Tang, B. Fang, Multi-focus image fusion based on the neighbor distance, Pattern Recogn. 46 (3) (2013) 1002–1011.
[14] N. Mitianoudis, T. Stathaki, Pixel-based and region-based image fusion schemes using ICA bases, Inf. Fus. 8 (2) (2007) 131–142.
[15] B. Yang, S. Li, Multifocus image fusion and restoration with sparse representation, IEEE Trans. Instrum. Measur. 59 (4) (2010) 884–892.
[16] J. Liang, Y. He, D. Liu, X. Zeng, Image fusion using higher order singular value decomposition, IEEE Trans. Image Process. 21 (5) (2012) 2898–2909.
[17] T. Wan, C. Zhu, Z. Qin, Multi-focus image fusion based on robust principal component analysis, Pattern Recogn. Lett. 34 (9) (2013) 1001–1008.
[18] S. Li, J. Kwok, Y. Wang, Combination of images with diverse focuses using the spatial frequency, Inf. Fus. 2 (3) (2001) 169–176.
[19] V. Aslantas, R. Kurban, Fusion of multi-focus images using differential evolution algorithm, Expert Syst. Appl. 37 (12) (2010) 8861–8870.
[20] M. Li, W. Cai, Z. Tan, A region-based multi-sensor image fusion scheme using pulse-coupled neural network, Pattern Recogn. Lett. 27 (16) (2006) 1948–1956.
[21] S. Li, B. Yang, Multifocus image fusion using region segmentation and spatial frequency, Image Vis. Comput. 26 (7) (2008) 971–979.
[22] W. Huang, Z. Jing, Evaluation of focus measures in multi-focus image fusion, Pattern Recogn. Lett. 28 (4) (2007) 493–500.
[23] S. Li, X. Kang, J. Hu, B. Yang, Image matting for fusion of multi-focus images in dynamic scenes, Inf. Fus. 14 (2) (2013) 147–162.
[24] S. Li, X. Kang, J. Hu, Image fusion with guided filtering, IEEE Trans. Image Process. 22 (7) (2013) 2864–2875.
[25] R. Eckhorn, H. Reitboeck, M. Arndt, P. Dicke, Feature linking via synchronization among distributed assemblies: simulation of results from cat cortex, Neural Comput. 2 (3) (1990) 293–307.
[26] R. Broussard, S. Rogers, M. Oxley, G. Tarr, Physiologically motivated image fusion for object detection using a pulse coupled neural network, IEEE Trans. Neural Networks 10 (3) (1999) 554–563.
[27] W. Huang, Z. Jing, Multi-focus image fusion using pulse coupled neural network, Pattern Recogn. Lett. 28 (9) (2007) 1123–1132.
[28] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[29] C. Liu, J. Yuen, A. Torralba, J. Sivic, W. Freeman, Sift flow: dense correspondence across different scenes, in: Proceedings of 10th European Conference on Computer Vision, 2008, pp. 28–42.
[30] 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.
[31] C. Liu, 2009. <http://people.csail.mit.edu/celiu/ECCV2008/>.
[32] R. Gonzalez, R. Woods, S. Eddins, Digital Image Processing Using MATLAB, 2nd edition., Gatesmark Publishing, 2009.
[33] A. Eskicioglu, P.S. Fisher, Image quality measures and their performance, IEEE Trans. Commun. 43 (12) (1995) 2959–2965.
[34] O. Rockinger, 1999. <http://www.metapix.de/toolbox.htm>.
[35] Mathworks, 2013. <http://www.mathworks.cn/matlabcentral/>.
[36] X. Qu, 2012. <http://www.quxiaobo.org/index.html>.
[37] X. Kang, 2013. <http://xudongkang.weebly.com/index.html>.
[38] 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.
[39] M. Hossny, S. Nahavandi, D. Creighton, Comments on information measure for performance of image fusion, Electron. Lett. 44 (18) (2008) 1066–1067.
[40] G. Qu, D. Zhang, P. Yan, Information measure for performance of image fusion, Electron. Lett. 38 (7) (2002) 313–315.
[41] Q. Wang, Y. Shen, A nonlinear correlation measure for multivariable data set, Phys. D: Nonlinear Phenom. 200 (3) (2005) 287–295.
[42] C.S. Xydeas, V.S. Petrovic, Objective image fusion performance measure, Electron. Lett. 36 (4) (2000) 308–309.
[43] J. Zhao, R. Laganiere, Z. Liu, Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement, Int. J. Innovat. Comput. Inf. Control 6 (A3) (2007) 1433–1447.
[44] C. Yang, J. Zhang, X. Wang, X. Liu, A novel similarity based quality metric for image fusion, Inf. Fus. 9 (2) (2008) 156–160.
[45] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[46] G. Piella, H. Heijmans, A new quality metric for image fusion, in: Proceedings of 10th International Conference on Image Processing, 2003, pp. 173–176.
[47] Y. Chen, R. Blum, A new automated quality assessment algorithm for image fusion, Image Vis. Comput. 27 (10) (2009) 1421–1432.
[48] http://www.wetcanvas.com/ArtSchool/Hagan/depthofield.htm, 1999.
[49] http://www.ucassdl.cn/resource.asp, 2009.
[50] Z. Zhang, R. Blum, A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application, Proc. IEEE 87 (8) (1999) 1315–1326.