Information Fusion 24 (2015) 147–164

Contents lists available at ScienceDirect

Information Fusion

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

A general framework for image fusion based on multi-scale transform and sparse representation

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 28 January 2014
Received in revised form 1 September 2014
Accepted 9 September 2014
Available online 5 October 2014

Keywords:
Image fusion
Multi-scale transform
Sparse representation

ABSTRACT

In image fusion literature, multi-scale transform (MST) and sparse representation (SR) are two most widely used signal/image representation theories. This paper presents a general image fusion framework by combining MST and SR to simultaneously overcome the inherent defects of both the MST- and SR-based fusion methods. In our fusion framework, the MST is firstly performed on each of the pre-registered source images to obtain their low-pass and high-pass coefficients. Then, the low-pass bands are merged with a SR-based fusion approach while the high-pass bands are fused using the absolute values of coefficients as activity level measurement. The fused image is finally obtained by performing the inverse MST on the merged coefficients. The advantages of the proposed fusion framework over individual MST- or SR-based method are first exhibited in detail from a theoretical point of view, and then experimentally verified with multi-focus, visible-infrared and medical image fusion. In particular, six popular multi-scale transforms, which are Laplacian pyramid (LP), ratio of low-pass pyramid (RP), discrete wavelet transform (DWT), dual-tree complex wavelet transform (DTCWT), curvelet transform (CVT) and nonsubsampled contourlet transform (NSCT), with different decomposition levels ranging from one to four are tested in our experiments. By comparing the fused results subjectively and objectively, we give the best-performed fusion method under the proposed framework for each category of image fusion. The effect of the sliding window's step length is also investigated. Furthermore, experimental results demonstrate that the proposed fusion framework can obtain state-of-the-art performance, especially for the fusion of multimodal images.

1. Introduction

In recent years, image fusion has become an important issue in image processing community. The target of image fusion is to generate a composite image by integrating the complementary information from multiple source images of the same scene [1]. For an image fusion system, the input source images can be acquired from either different types of imaging sensors or a sensor whose optical parameters can be changed, and the output called fused image will be more suitable for human or machine perception than any individual source image. Image fusion technique has been widely employed in many applications such as computer vision, surveillance, medical imaging, and remote sensing.

Multi-scale transform (MST) theories are the most popular tools used in various image fusion scenarios such as multi-focus image fusion, visible-infrared image fusion, and multimodal medical image fusion. Classical MST-based fusion methods include pyramid-based ones like Laplacian pyramid (LP) [2], ratio of low-pass pyramid (RP) [3] and gradient pyramid (GP) [4], wavelet-based ones like discrete wavelet transform (DWT) [5], stationary wavelet transform (SWT) [6] and dual-tree complex wavelet transform (DTCWT) [7], and multi-scale geometric analysis (MGA)-based ones like curvelet transform (CVT) [8] and nonsubsampled contourlet transform (NSCT) [9]. In general, the MST-based fusion methods consist of the following three steps [10]. First, decompose the source images into a multi-scale transform domain. Then, merge the transformed coefficients with a given fusion rule. Finally, reconstruct the fused image by performing the corresponding inverse transform over the merged coefficients. These methods assume that the underlying salient information of the source images can be extracted from the decomposed coefficients. Obviously, the selection of transform domain plays a crucial role

* 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.2014.09.004
1566-2535/

in these methods. A comparative study of different MST-based methods is reported in [11], where Li et al. found that the NSCT-based method can generally achieve the best results. In addition to the selection of transform domain, the fusion rule in either high-pass or low-pass band also has a great impact on the fused results. Conventionally, the absolute value of high-pass coefficient is used as the activity level measurement for high-pass fusion. The simplest rule is selecting the coefficient with largest absolute value at each pixel position (the "max-absolute" rule). Many improved high-pass fusion rules which make use of the neighbor coefficients' information have also been developed. However, compared with the great concentration on developing effective rules for high-pass fusion, less attention has been paid to the fusion of low-pass bands. In most MST-based fusion methods, the low-pass bands are just simply merged by averaging all the source inputs (the "averaging" rule).

Sparse representation addresses the signals' natural sparsity, which is in accord with the physiological characteristics of human visual system [12]. The basic assumption behind SR is that a signal $\mathbf{x} \in \mathbf{R}^n$ can be approximately represented by a linear combination of a "few" atoms from an overcomplete dictionary $\mathbf{D} \in \mathbf{R}^{n\times m}(n < m)$, where n is the signal dimension and m is the dictionary size. That is, the signal $\mathbf{x}$ can be expressed as $\mathbf{x} \approx \mathbf{D}\alpha$, where $\alpha \in \mathbf{R}^m$ is the unknown sparse coefficient vector. As the dictionary is overcomplete, there are numerous feasible solutions for this underdetermined system. The target of SR is to calculate the sparsest α which contains the fewest nonzero entries among all feasible solutions (known as sparse coding). In SR-based image processing methods, the sparse coding technique is often performed on local image patches for the sake of algorithm stability and efficiency [13]. Yang and Li [14] first introduced SR into image fusion. The sliding window technique (patches are overlapped) is adopted in their method to make the fusion process more robust to noise and misregistration. In [14], the sparse coefficient vector is used as the activity level measurement. Particularly, among all the source sparse vectors, the one owning the maximal $l_1 - norm$ is selected as the fused sparse vector (the "max-L1" rule). The fused image is finally reconstructed with all the fused sparse vectors. Their experimental results show that the SR-based fusion method owns clear advantages over traditional MST-based methods for multi-focus image fusion, and can lead to state-of-the-art results. In the past few years, the SR-based fusion has emerged as a new active branch in image fusion research with many improved approaches being proposed [15–18].

Although both the MST- and SR-based methods have achieved great success in image fusion, it is worthwhile to notice that both of them have some defects, which will be further discussed in this paper. Moreover, to overcome the related disadvantages, we present a general image fusion framework by taking the complementary advantages of MST and SR. Specifically, the low-pass MST bands are merged with a SR-based fusion approach while the high-pass MST bands are fused using the conventional "max-absolute" rule with a local window based consistency verification scheme [5]. To verify the effectiveness of the proposed framework, six popular multi-scale transforms (MSTs), which are LP, RP, DWT, DTCWT, CVT and NSCT, with different decomposition levels ranging from one to four are tested in our experiments. By comparing the fused results subjectively and objectively, we give the best-performed methods under the proposed framework for the fusion of multi-focus, visible-infrared and medical images, respectively. The effect of the sliding window's step length is also investigated. Experimental results demonstrate that the combined methods can clearly outperform both the MST- and SR-based methods. Furthermore, the proposed fusion methods can obtain state-of-the-art fused results, especially for the fusion of medical images as well as visible-infrared images.

The rest of this paper is organized as follows. We first present the detailed fusion framework in Section 2. In Section 3, the disadvantages of MST- and SR-based methods and why the proposed framework can overcome them are discussed from a theoretical perspective. The experimental results are given in Section 4. Section 5 summarizes some main conclusions of this paper.

2. Proposed fusion framework

To better exhibit the advantages of the proposed framework over MST- and SR-based methods, we first present the details of our framework in this section.

2.1. Dictionary learning

The overcomplete dictionary determines the signal representation ability of sparse coding. Generally, there are two main categories of offline approaches to obtain a dictionary. The first one is directly using the analytical models such as discrete cosine transform (DCT) and CVT. However, this category of dictionary is restricted to signals of a certain type and cannot be used for an arbitrary family of signals. The second category is applying the machine learning technique to obtain the dictionary from a large number of training image patches. Suppose that M training patches of size $\sqrt{n} \times \sqrt{n}$ are rearranged to column vectors in the $\mathbf{R}^n$ space, thereby the training database $\{\mathbf{y}_i\}_{i=1}^M$ is constructed with each $\mathbf{y}_i \in \mathbf{R}^n$. The dictionary learning model can be presented as

$$\min_{\mathbf{D},\{\alpha_i\}_{i=1}^M} \sum_{i=1}^{M} \|\alpha_i\|_0 \quad s.t. \quad \|\mathbf{y}_i - \mathbf{D}\alpha_i\|_2 < \varepsilon, \quad i \in \{1,\ldots,M\}, \tag{1}$$

where $\varepsilon > 0$ is an error tolerance, $\{\alpha_i\}_{i=1}^M$ is the unknown sparse vectors corresponding to $\{\mathbf{y}_i\}_{i=1}^M$ and $\mathbf{D} \in \mathbf{R}^{n\times m}$ is the unknown dictionary to be learned. Some effective methods such as MOD [19] and K-SVD [20] have been proposed to solve this problem. The learned dictionaries usually have better representation ability than the pre-constructed ones, so we adopt the learning-based approach in this paper.

In this work, the sparse coding technique is employed for the fusion of MST low-pass bands. One possible way to get the training patches is sampling from the corresponding MST low-pass bands which are obtained from some training images under the same decomposition condition. However, in this case, the dictionary learning process should be repeated once either the selected transform domain or even one specific parameter (such as the decomposition level or selected image filter) is changed. Obviously, this will decrease the flexibility and practicality of the fusion method to a large extent. In this paper, we aim to learn a universal dictionary which can be used in any specific transform domain and parameter settings. As is well known, the MST low-pass band obtained by image filtering can be viewed as a smooth version of the original image. Since the numerous "flat" patches contained in a natural image can be well sparsely represented by a dictionary learned from natural image patches, it is theoretically feasible to use the same dictionary to represent the patches in the low-pass bands so long as the mean value of each sampled patch is subtracted to zero before training. In this situation, the mean value of each atom in the obtained dictionary is also zero, so the atoms only contain structural information. For an input patch to be represented, its mean value should also be subtracted to zero before sparse coding. Thus, we can directly use natural image patches to learn a universal dictionary.

2.2. Detailed fusion scheme

The schematic diagram of the proposed fusion framework is shown in Fig. 1. For simplicity, only the fusion of two source images is considered while the proposed framework can be straightforwardly extended to fuse more than two images. The detailed fusion scheme contains the following four steps.

Step 1: MST decomposition.
Perform a specific MST on the two source images $\{I_A, I_B\}$ to obtain their low-pass bands $\{L_A, L_B\}$ and high-pass bands which are uniformly denoted as $\{H_A, H_B\}$.

Step 2: Low-pass fusion.

(i) Apply the sliding window technique to divide L_A and L_B into image patches of size $\sqrt{n} \times \sqrt{n}$ from upper left to lower right with a step length of s pixels. Suppose that there are T patches denoted as $\{p_A^i\}_{i=1}^T$ and $\{p_B^i\}_{i=1}^T$ in L_A and L_B, respectively.

(ii) For each position i, rearrange $\{p_A^i, p_B^i\}$ into column vectors $\{\mathbf{v}_A^i, \mathbf{v}_B^i\}$ and then normalize each vector's mean value to zero to obtain $\{\hat{\mathbf{v}}_A^i, \hat{\mathbf{v}}_B^i\}$ by

$$\hat{\mathbf{v}}_A^i = \mathbf{v}_A^i - \bar{v}_A^i \cdot \mathbf{1}, \tag{2}$$

$$\hat{\mathbf{v}}_B^i = \mathbf{v}_B^i - \bar{v}_B^i \cdot \mathbf{1}, \tag{3}$$

where $\mathbf{1}$ denotes an all-one valued $n \times 1$ vector, $\bar{v}_A^i$ and $\bar{v}_B^i$ are the mean values of all the elements in $\mathbf{v}_A^i$ and $\mathbf{v}_B^i$, respectively.

(iii) Calculate the sparse coefficient vectors $\{\alpha_A^i, \alpha_B^i\}$ of $\{\hat{\mathbf{v}}_A^i, \hat{\mathbf{v}}_B^i\}$ using the orthogonal matching pursuit (OMP) algorithm [21] by

$$\alpha_A^i = \arg\min_{\alpha} ||\alpha||_0 \quad s.t. \quad ||\hat{\mathbf{v}}_A^i - \mathbf{D}\alpha||_2 < \varepsilon, \tag{4}$$

$$\alpha_B^i = \arg\min_{\alpha} ||\alpha||_0 \quad s.t. \quad ||\hat{\mathbf{v}}_B^i - \mathbf{D}\alpha||_2 < \varepsilon, \tag{5}$$

where $\mathbf{D}$ is the learned dictionary.

(iv) Merge α_A^i and α_B^i with the "max-L1" rule to obtain the fused sparse vector

$$\alpha_F^i = \begin{cases} \alpha_A^i & if \quad \|\alpha_A^i\|_1 > \|\alpha_B^i\|_1 \\ \alpha_B^i & \text{otherwise} \end{cases}. \tag{6}$$

The fused result of $\mathbf{v}_A^i$ and $\mathbf{v}_B^i$ is calculated by

$$\mathbf{v}_F^i = \mathbf{D}\alpha_F^i + \bar{v}_F^i \cdot \mathbf{1}, \tag{7}$$

where the merged mean value $\bar{v}_F^i$ is obtained by

$$\bar{v}_F^i = \begin{cases} \bar{v}_A^i & if \quad \alpha_F^i = \alpha_A^i \\ \bar{v}_B^i & \text{otherwise} \end{cases}. \tag{8}$$

(v) Iterate the above process for all the source image patches in $\{p_A^i\}_{i=1}^T$ and $\{p_B^i\}_{i=1}^T$ to obtain all the fused vectors $\{\mathbf{v}_F^i\}_{i=1}^T$. Let L_F denotes the low-pass fused result. For each $\mathbf{v}_F^i$, reshape it into a patch p_F^i and then plug p_F^i into its original position in L_F. As patches are overlapped, each pixel's value in L_F is averaged over its accumulation times.

Step 3: High-pass fusion.
Merge H_A and H_B to obtain H_F with the popular "max-absolute" rule using the absolute value of each coefficient as the activity level measurement. Then, apply the consistency verification scheme (see in [5]) to ensure that a fused coefficient does not originate from a different source image from most of its neighbors. This can be implemented via a small majority filter.

Step 4: MST reconstruction.
Perform the corresponding inverse MST over L_F and H_F to reconstruct the final fused image I_F.

3. Why the proposed framework works

In this section, for each of the MST- and SR-based fusion methods, we first itemize its main defects and then show why

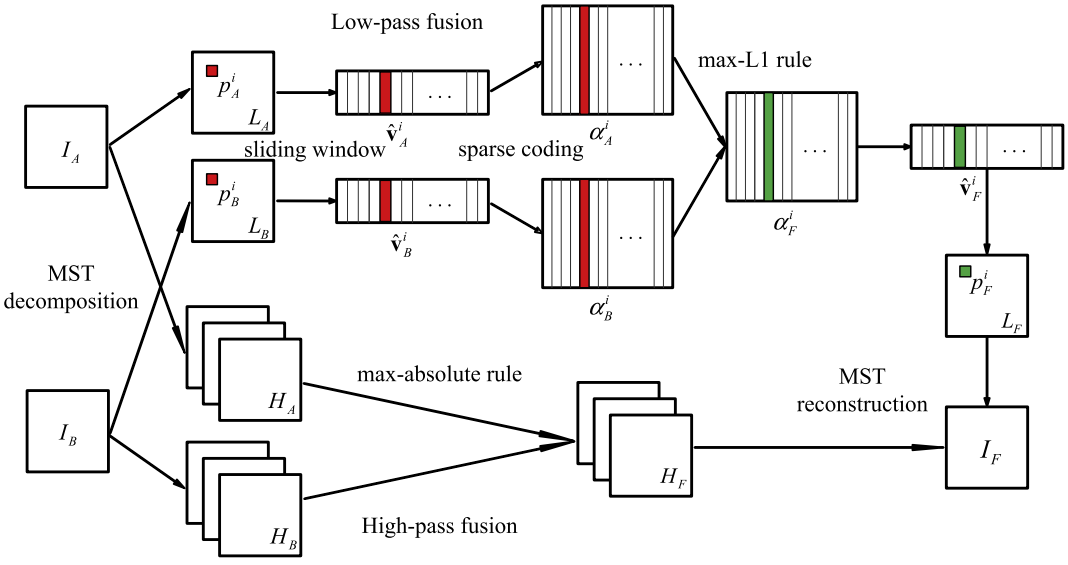


Fig. 1. The schematic diagram of the proposed fusion framework.

the proposed framework can overcome them. All the points given here will be further experimentally verified in Section 4.

3.1. Advantages over the MST-based methods

For the conventional MST-based image fusion methods (the high-pass bands are merged with the "max-absolute" rule while the low-pass bands are fused using the "averaging" rule), there are two main drawbacks as follows.

The first one is the loss of contrast. Since most energy of an image is contained in the low-pass band (even though the decomposition level is set to 4 according to the analysis in [11]), the "averaging" fusion rule tends to lose some energy in the source images. For multi-focus image fusion, this phenomenon is not obvious because the source images are captured from the same type of sensors. However, for the fusion of multimodal images such as visible-infrared and medical images, the fused results of the MST-based methods are often in low contrast. This is mainly because different imaging modalities reflect different physical attributes, so a same region in different source images may have different brightness. For example, Fig. 2(a) shows a pair of computed tomography (CT) and magnetic resonance (MR) images. It can be seen that the CT image mainly focuses on dense structures like bones, while the MR image provides excellent soft-tissue details. When the "averaging" rule is used for low-pass fusion, the energy contained in those regions will be lost to a large extent. As a result, the contrast of those regions in the fused image will decrease a lot after MST reconstruction.

The second one is the difficulty in selecting the MST decomposition level. On one hand, to ensure enough spatial details can be extracted from the source images, the decomposition level cannot be too small such as 1 or 2. On the other hand, Li et al. [11] experimentally verified that when the decomposition level is too large, one coefficient in the low-pass band have an impact on a large set of pixels in the fused image, so an error in the low-pass band (mainly caused by noise or mis-registration between the source images) will lead to serious artificial effects. Moreover, when the decomposition level becomes larger, the quality of high-pass fusion is also more sensitive to noise and mis-registration. Therefore, when the source images are not precisely registered, the decomposition level cannot be too large. Particularly, for multi-focus image fusion, due to the different imaging parameters (e.g. focal length) for multiple source images, the locations of object edges in different source images are often not exactly the same for their different sharpness. A typical example is shown in Fig. 2(b). Between the two source images, both the borders and numbers of the two clocks in the scene have different sharpness, so it is practically impossible to make an accurate registration. Thus, a compromise on decomposition level should be made for the consideration of extracting enough spatial details and being robust to mis-registration. Although a recommended value of 4 is given in [11], we experimentally find that the MST-based methods are still sensitive to mis-registration (results are shown in Section 4).

As a smart blending approach, the SR-based image fusion scheme is combined into the MST-based fusion methods to overcome the above two defects. In the proposed framework, the SR-based scheme is employed to fuse the MST low-pass bands. In Section 2, after applying the "max-L1" rule in Eq. (6), we transfer the energy in source images to the fused image by Eq. (8). Therefore, the contrast in the fused image is improved. For the second defect, by extracting spatial details in low-pass band with the SR-based fusion scheme, the decomposition level can be set less than 4 for multi-focus image fusion to make the method more robust to mis-registration. Thus, the difficulty in determining decomposition level can be well solved.

3.2. Advantages over the SR-based method

The conventional SR-based image fusion method [14] mainly has the following three defects.

The first one is the fine details in source images like textures and edges tend to be smoothed for the following two reasons. First, the signal representation ability of the dictionary may be not sufficient for fine details, which means that the reconstruction result is not approximate to the input signal. As we know, the representation ability of the over-completed dictionary relies much on the number of atoms in it, but a dictionary with a large size will directly increase the computational cost. More importantly, the study in [22] shows that a highly redundant dictionary may lead to potential visual artifacts in the reconstruction result, especially when the input signal is corrupted by noise. Thus, a compromise on dictionary size is usually required. A typical example is that the dictionary size is 256 when the input signal is 64 dimensional (8×8 image patch) [15]. Second, the usage of sliding window technique may also cause smoothness. The step length of the sliding window is usually set to 1 when fusing images directly in spatial domain to avoid blocking effects [14]. However, when the adjacent patches are greatly overlapped, some details in the fused image will be smoothed.

The second one is the "max-L1" rule may cause spatial inconsistency in the fused image when the source images are captured by different imaging modalities. As mentioned before, for multimodal image fusion, a region may be very bright in one source image while very dark in another, but the region in both of them may be very "flat" with few fine details. Note that although a region in each of two source images is visually "flat", there still exists little difference between the two source images in terms of variance, and the difference is usually consistent over all the patches in that region. That is to say, if one patch in the region of source image *A* has a larger variance than the corresponding patch in source image *B*, then most of the other patches in that region of source image *A* also tend to have larger variances than the corresponding patches

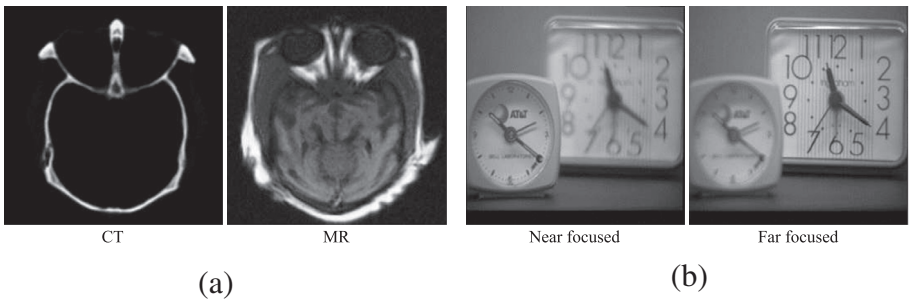


Fig. 2. Two pairs of source images. (a) Medical images, and (b) multi-focus images.

in source image B. However, since the difference is very tiny, the "max-L1" fusion rule will become very sensitive to the random noise in spatial domain because a small change of value at a pixel may influence the fusion result of several patches. As a result, the fused patches in that region may originate from different source images, which will lead to spatial inconsistency in the fused image. Since the SR-based method handles patches in spatial domain, the impact of high-frequency noise is considerable.

The third one is the low computational efficiency. Since the sliding window's step length should be small enough, the sparse coding technique is performed on a large number of image patches. For instance, when the patch size is 8×8 and the step length is set to 1, there are 62001 patches to be processed for a source image of size 256×256. In this case, it usually takes several minutes to fuse two source images with the SR-based method.

The proposed fusion framework can effectively overcome the above three defects of the SR-based method. In our fusion framework, the high-frequency spatial information is separated by performing MST and extracted by the "max-absolute" rule. Meanwhile, the representation ability of the dictionary is enough to satisfy the reconstruction accuracy for low-frequency components. Furthermore, we will show in the next section that the sliding window's step length in low-pass bands can be set larger than that in spatial domain. Therefore, the inclination of SR-based method to smooth fine details can be prevented. For the second defect, without high-frequency details, the random noise can be effectively eliminated, so the probability that the patches in a "flat" region originate from different source images will decrease to a large extent, leading to better spatial consistency. Finally, the computational efficiency can also be improved by the proposed framework because the number of patches required to be processed with the sparse coding technique is greatly reduced. For one thing, the step length can be set larger. For another, the low-pass bands of many MSTs such as LP and DWT have smaller size relative to the original image.

4. Experiments

4.1. Experimental setups

4.1.1. Source images

As shown in Fig. 3, 26 pairs of source images grouped into three categories are employed to verify the effectiveness of the proposed fusion framework. Among them, there are 10 pairs of multi-focus images (Figs. 3(a)), 8 pairs of visible-infrared images (Fig. 3(b)) and 8 pairs of medical images (Fig. 3(c)). For each pair, the two source images are assumed to be pre-registered in our study.

4.1.2. Objective evaluation metrics

It is not an easy task to quantitatively evaluate the quality of a fused image since the reference image (ground truth) does not exist in practice. In recent years, many fusion metrics have been proposed, but none of them is universally believed to be always more reasonable than others for various fusion scenarios. Thus, it is usually necessary to apply several metrics to make a comprehensive evaluation. In this work, five popular metrics, which are briefly introduced as follows, are employed to quantitatively evaluate the performances of different fusion methods. Uniformly, let A and B denote two source images of size $H \times W$ while F represents the fused image.

1. Standard deviation (SD). The SD of the fused image is defined as

$$SD = \sqrt{\frac{1}{H \times W}\sum_{x=1}^{H}\sum_{y=1}^{W}(F(x,y)-\mu)^2}, \tag{9}$$

where μ is the mean value of the fused image. SD is mainly used to measure the overall contrast of the fused image.

2. Entropy (EN). The EN of the fused image is defined as

$$EN = -\sum_{l=0}^{L-1} p_F(l)\log_2 p_F(l), \tag{10}$$

where L is the number of gray level and $p_F(l)$ is the normalized histogram of the fused image. In our experiments, L is set to 256. EN is used to measure the amount of information in the fused image.

3. The gradient based fusion metric Q_G proposed by Xydeas and Petrovic [23]. It 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 $Q^{AF}(x,y) = Q_g^{AF}(x,y)Q_\alpha^{AF}(x,y)$, $Q_g^{AF}(x,y)$ 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 the same as that of $Q^{AF}(x,y)$. 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. The Q_G is a popular fusion metric which computes the amount of gradient information injected into the fused image from the source images.

4. The phase congruency based fusion metric Q_P proposed by Zhao et al. [24]. It is based on the principal moments of the image phase congruency, which reflect 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 refer to phase congruency, maximum and minimum moments, respectively. The exponential parameters α, β and γ are all set to 1 in this work. More details about this metric can be found in [24]. The Q_P measures the extent that the salient features in the source images are preserved.

5. The universal image quality index (UIQI) [25] based fusion metric Q_W proposed by Piella and Heijmans [26]. The Q_W is defined as

$$Q_W = \sum_{w\in W} c(w)(\lambda(w)Q_0(A,F|w) + (1 - \lambda(w))Q_0(B,F|w)), \tag{13}$$

where $Q_0(A,F|w)$ and $Q_0(B,F|w)$ are calculated using the method in [25] in a local sliding window w. The saliency weight $\lambda(w)$ is calculated by

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

where the salience measure $s(A|w)$ and $s(B|w)$ are calculated with the variance of A and B in window w, respectively. The $c(w)$ is the normalized salience of w among all the local windows. The $c(w)$ is obtained by

$$c(w) = \frac{\max(s(A|w), s(B|w))}{\sum_{w'\in W}\max(s(A|w'), s(B|w'))}. \tag{15}$$

Based on the UIQI [25], the metric Q_W firstly addresses the distortions of coefficient correlation, illumination and contrast between source images and the fused image, which are in accord with the characteristics of human visual system. In addition, it also takes image salience into consideration.

For each of the five metrics, a larger value generally indicates a better fused result. To guarantee the objectivity of evaluation results, each of Q_G, Q_P and Q_W is calculated with a widely used implementation from a third party. Specifically, the code of Q_G is available on website [27] implemented by Qu. The code of Q_P is obtained from a fusion evaluation toolbox [28] provided

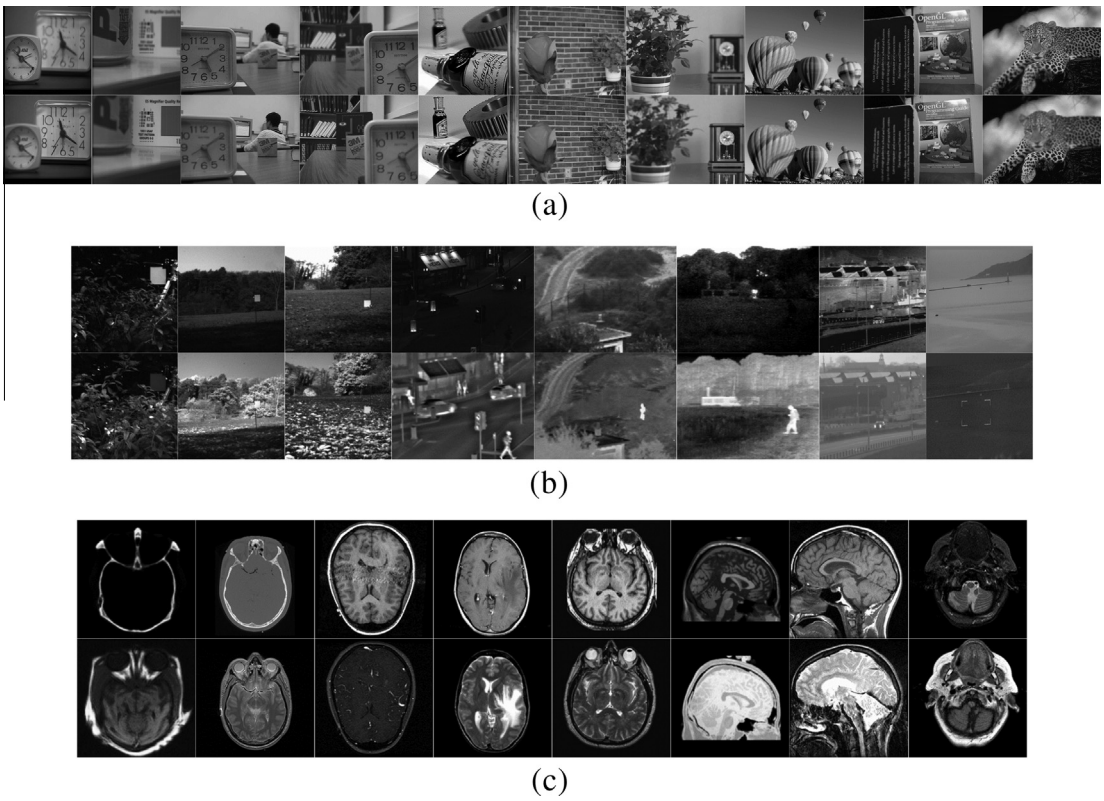


Fig. 3. The source images used in our experiments. (a) Multi-focus images (10 pairs), (b) visible-infrared images (8 pairs), and (c) medical images (8 pairs).

by Liu (the third author of [24]). The code of Q_W is provided by Kang and Li (the first author of [11]). Furthermore, all the parameters in the above metrics are set to the default values reported in the related publications.

4.2. Experimental results on six popular MSTs

In this subsection, six popular MSTs, which are Laplacian pyramid (LP), ratio of low-pass pyramid (RP), discrete wavelet transform (DWT), dual-tree complex wavelet transform (DTCWT), curvelet transform (CVT) and nonsubsampled contourlet transform (NSCT), are used to exhibit the advantages of the proposed fusion framework over the MST- and SR-based methods. First, for each specific MST, the effectiveness of the combined method is verified using the above objective fusion metics, and the impact of the decomposition level is studied by increasing from 1 to 4. Then, we make an overall comparison in terms of both objective assessment and subjective visual quality. Particularly, for each of the three categories of image fusion, we give the best-performed method, which includes a specific MST and its decomposition level, under the proposed framework. For convenience, a fusion method under the proposed framework using a certain MST with a decomposition level of L is denoted as MST-SR-L. For example, DWT-SR-3 represents the method by combining DWT and SR with a 3-level decomposition. Moreover, for each MST-based method, the decomposition level is just set to 4 according to the analysis in [11]. For all the MST- and MST-SR-based methods, the "max-absolute" rule with a 3×3 window based consistency verification scheme [5] is adopted to merge the high-pass bands.

It is worthwhile to notice that the image filters (if any) used in MSTs are fixed in our experiments for the following two reasons. First, according to [11], the impact of image filters on the fused results is generally much weaker than that of either transform domain or decomposition level. Second, it is very difficult to give a comprehensive investigation since there may be many selectable filters for a MST, leading to numerous parameter combinations. Therefore, in this work, we fix its image filters mainly based on the related results reported in [11]. The detailed parameter settings of each MST will be given separately later. For every fusion method which uses the sparse coding model, the local patch size is set to 8×8 and the error tolerance ε is set to 0.1 according to the analysis in [14]. The step length of the sliding window in this subsection is set to 1. The dictionary used in the sparse model is learned by the K-SVD method [20]. The training data consists of 100,000 8×8 patches, which are randomly sampled from a database of 40 high-quality natural images. The dictionary size is set to 256 and the iteration number of K-SVD is fixed to 180.

4.2.1. LP-SR

In all the following tables in this paper, each reported value is the average result of all the source images in the same category, and a value labeled in bold indicates the best performance over all the methods on the corresponding fusion metric. Table 1 lists the objective assessment of the LP, SR and LP-SR methods. We can see from Table 1 that for multi-focus image fusion, when the decomposition level becomes larger, both *SD* and *EN* increase, but Q_G, Q_P and Q_W generally decrease. For either visible-infrared or medical image fusion, the best performances on all the five metrics except for *EN* in visible-infrared image fusion come from the LP-SR-4 method. Furthermore, the advantage of the LP-SR-4 method over either the LP or SR method is very clear. In general, a higher decomposition level leads to a better fusion performance.

4.2.2. RP-SR

The objective assessment of the RP, SR and RP-SR methods is listed in Table 2. For multi-focus image fusion, the situation is very similar to that in Table 1. The influence of decomposition level on the five metrics are very clear. However, for either visible-infrared

Table 1
Objective assessment of the LP, SR and LP-SR methods.

Images	Metrics	LP	SR	LP-SR-1	LP-SR-2	LP-SR-3	LP-SR-4
Multi-focus	*SD*	52.3502	51.9298	52.2052	52.2727	52.3038	**52.3657**
	EN	7.3085	7.2997	7.3024	7.3054	7.3086	**7.3101**
	Q_G	0.7578	**0.7666**	0.7650	0.7624	0.7600	0.7599
	Q_P	0.9022	0.9064	**0.9095**	0.9057	0.9035	0.9025
	Q_W	0.9249	**0.9264**	0.9253	0.9250	0.9251	0.9249
Visible-infrared	*SD*	39.1052	40.5449	43.0950	43.6619	44.7266	**45.1038**
	EN	6.7938	6.8404	6.9619	7.0449	**7.1434**	7.1237
	Q_G	0.6462	0.6020	0.6433	0.6514	0.6596	**0.6698**
	Q_P	0.5175	0.4166	0.4981	0.5284	0.5412	**0.5514**
	Q_W	0.7659	0.6649	0.7438	0.7844	0.7948	**0.7960**
Medical	*SD*	64.4253	65.1178	65.0250	66.5099	69.8084	**72.7315**
	EN	5.5393	5.7777	5.8087	5.8007	5.8249	**5.8372**
	Q_G	0.6258	0.5992	0.6218	0.6337	0.6405	**0.6459**
	Q_P	0.4613	0.4019	0.4356	0.4666	0.4774	**0.4905**
	Q_W	0.7678	0.7504	0.7659	0.7705	0.7862	**0.7996**

Table 2
Objective assessments of the RP, SR and RP-SR methods.

Images	Metrics	RP	SR	RP-SR-1	RP-SR-2	RP-SR-3	RP-SR-4
Multi-focus	*SD*	52.1889	51.9298	52.0818	52.1169	52.1663	**52.2109**
	EN	7.3079	7.2997	7.3020	7.3069	7.3089	**7.3107**
	Q_G	0.7566	**0.7666**	0.7645	0.7599	0.7570	0.7568
	Q_P	0.8855	**0.9064**	0.8962	0.8885	0.8864	0.8857
	Q_W	0.9224	**0.9264**	0.9257	0.9229	0.9228	0.9224
Visible-infrared	*SD*	38.6841	40.5449	43.5556	44.9450	**45.7618**	45.0224
	EN	6.7511	6.8404	6.9517	7.0577	**7.1423**	7.1170
	Q_G	0.4820	**0.6020**	0.5366	0.5040	0.4896	0.4782
	Q_P	0.3803	**0.4166**	0.3501	0.3507	0.3727	0.3763
	Q_W	0.5333	0.6649	**0.6910**	0.6769	0.6424	0.6066
Medical	*SD*	63.1476	65.1178	65.2303	67.1031	70.8526	**74.4538**
	EN	5.7861	5.7777	5.8100	5.8491	5.8637	**5.9122**
	Q_G	0.4384	**0.5992**	0.4947	0.4650	0.4554	0.4240
	Q_P	0.3567	**0.4019**	0.2756	0.2952	0.3372	0.3264
	Q_W	0.5607	**0.7504**	0.7098	0.6615	0.6366	0.5985

or medical image fusion, it can be seen from Table 2 that the values of Q_G, Q_P and Q_W are much lower than the corresponding ones in Table 1. Therefore, the RP and RP-SR methods may be not applicable for the fusion of multimodal images.

4.2.3. DWT-SR

Table 3 gives the objective assessment of the DWT, SR and DWT-SR methods. The wavelet basis 'db1' is applied in all the DWT and DWT-SR methods [11]. For multi-focus image fusion, the situation is also similar to that of the LP-based results listed in Table 1, while a main difference is that the DWT-SR-1 method outperforms the SR method on both Q_G and Q_P. The situation of either visible-infrared or medical image fusion is exactly the same as that in Table 1.

4.2.4. DTCWT-SR

The objective assessment of the DTCWT, SR and DTCWT-SR methods is given in Table 4. In the DTCWT and DTCWT-SR

Table 3
Objective assessment of the DWT, SR and DWT-SR methods.

Images	Metrics	DWT	SR	DWT-SR-1	DWT-SR-2	DWT-SR-3	DWT-SR-4
Multi-focus	*SD*	52.1126	51.9298	52.0612	52.0872	52.0264	**52.1284**
	EN	7.3405	7.2997	7.3121	7.3225	7.3300	**7.3424**
	Q_G	0.7349	0.7666	**0.7668**	0.7634	0.7512	0.7453
	Q_P	0.8723	0.9064	**0.9098**	0.9007	0.8792	0.8730
	Q_W	0.9227	**0.9264**	0.9259	0.9248	0.9238	0.9227
Visible-infrared	*SD*	36.5192	40.5449	42.2491	42.8619	43.4889	**44.1496**
	EN	6.7329	6.8404	6.9211	7.0487	**7.1284**	7.0965
	Q_G	0.5808	0.6020	0.6084	0.6045	0.6172	**0.6191**
	Q_P	0.4243	0.4166	0.4358	0.4398	0.4527	**0.4547**
	Q_W	0.7195	0.6649	0.7130	0.7385	0.7547	**0.7592**
Medical	*SD*	59.2863	65.1178	63.7895	63.2669	65.7815	**67.4361**
	EN	5.9469	5.7777	5.8087	5.8506	5.8936	**5.9854**
	Q_G	0.4921	0.5992	0.6039	0.6008	0.6112	**0.6193**
	Q_P	0.2836	0.4019	0.4041	0.4086	0.4258	**0.4437**
	Q_W	0.7098	0.7504	0.7492	0.7470	0.7535	**0.7627**

Table 4
Objective assessment of the DTCWT, SR and DTCWT-SR methods.

Images	Metrics	DTCWT	SR	DTCWT-SR-1	DTCWT-SR-2	DTCWT-SR-3	DTCWT-SR-4
Multi-focus	*SD*	52.0593	51.9298	52.0533	52.0337	52.0097	**52.0651**
	EN	7.3192	7.2997	7.3071	7.3103	7.3143	**7.3195**
	Q_G	0.7553	0.7666	**0.7688**	0.7642	0.7573	0.7577
	Q_P	0.9016	0.9064	**0.9114**	0.9081	0.9019	0.9019
	Q_W	0.9253	0.9264	0.9267	**0.9277**	0.9268	0.9254
Visible-infrared	*SD*	35.4250	40.5449	42.0276	44.0549	45.1063	**45.2558**
	EN	6.6795	6.8404	6.9054	7.0852	**7.1394**	7.1248
	Q_G	0.6260	0.6020	0.6440	0.6520	0.6603	**0.6738**
	Q_P	0.4911	0.4166	0.4544	0.4989	0.5396	**0.5646**
	Q_W	0.7209	0.6649	0.7071	0.7765	0.7870	**0.7940**
Medical	*SD*	58.5731	65.1178	64.6845	66.0199	67.7566	**69.8551**
	EN	5.8596	5.7777	5.8140	5.8470	5.9070	**5.9832**
	Q_G	0.5531	0.5992	0.6081	0.6037	0.6135	**0.6248**
	Q_P	0.3838	0.4019	0.4083	0.4265	0.4314	**0.4583**
	Q_W	0.7140	0.7504	0.7616	0.7617	0.7783	**0.7818**

methods, the image filters for the first-level and other-levels of decomposition are selected as LeGall 5-3 and Qshift-06 (quarter sample shift orthogonal 10-10 tap filter with 6-6 nonzero taps), respectively [11]. It can be seen from Table 4 that the DTCWT-SR-1 method can outperform the SR method on all the five metrics for the fusion of multi-focus images. For either visible-infrared or medical image fusion, the trend that a higher decomposition level can lead to a better fusion performance still exists.

4.2.5. CVT-SR

Table 5 lists the objective assessment of the CVT, SR and CVT-SR methods. It can be seen that the best performance on each metric is exactly the same as that of the LP-based results shown in Table 1 for all the three categories of image fusion.

4.2.6. NSCT-SR

The objective assessment of the NSCT, SR and NSCT-SR methods is given in Table 6. For the NSCT and NSCT-SR methods, we use the 'pyrexc' as the pyramid filter and the 'vk' as the directional filter [11]. Moreover, the direction numbers of the four decomposition levels from coarse to fine are selected as 4, 8, 8 and 16, respectively. We can see from Table 6 that the NSCT-SR-1 method gets the first place for Q_G, Q_P and Q_W with a clear advantage over the NSCT or SR method for multi-focus image fusion. For either visible-infrared or medical image fusion, just as most previous situations, the NSCT-SR-4 method clearly achieves the best performance over all the fusion methods.

4.2.7. Overall comparison

At last, for each type of image fusion, we take the related contents in Tables 1–6 into consideration together and seek out some common regularities among the six MSTs used in the proposed framework.

The characteristics of multi-focus image fusion can be summarized as the following three points. First, the MST methods can obtain almost the highest *SD* and *EN* while the lowest Q_G, Q_P and Q_W, which indicates that although they can extract enough spatial details with 4-level decomposition, they are very sensitive to mis-registration. Thus, the second defect of the MST methods mentioned in Section 3 is confirmed. Second, just on the contrary, the SR method achieves almost the largest Q_G, Q_P and Q_W while the lowest *SD* and *EN*, which means that it is robust to mis-registration but may tend to lose fine details, so the first defect of the SR method is at least partially verified. Third, the MST-SR methods can make a balance between the robustness to mis-registration and the ability of extracting spatial details. When the decomposition level becomes larger, both *SD* and *EN* increase while Q_G, Q_P and Q_W decrease. Moreover, it can be seen that the performance of the MST-SR-1 method is approximate to that of the SR method, while the performance of the MST-SR-4 method is approximate to that of the MST methods. For multi-focus image fusion, compared with the slight loss of spatial details, the robustness to mis-registration is more important to the visual quality of the fused image. Thus, the evaluations on Q_G, Q_P and Q_W are more meaningful to some extent. Based on this consideration, we can find that some

Table 5
Objective assessment of the CVT, SR and CVT-SR methods.

Images	Metrics	CVT	SR	CVT-SR-1	CVT-SR-2	CVT-SR-3	CVT-SR-4
Multi-focus	*SD*	52.0710	51.9298	52.0619	52.0541	52.0517	**52.0838**
	EN	7.3336	7.2997	7.3155	7.3247	7.3302	**7.3348**
	Q_G	0.7425	**0.7666**	0.7630	0.7546	0.7473	0.7429
	Q_P	0.8889	0.9064	**0.9086**	0.9029	0.8959	0.8891
	Q_W	0.9254	**0.9264**	0.9259	0.9255	0.9251	0.9252
Visible-infrared	*SD*	36.1982	40.5449	42.7476	43.8822	44.5739	**44.6091**
	EN	6.7154	6.8404	6.9275	7.0430	**7.1505**	7.1132
	Q_G	0.5903	0.6020	0.6017	0.6062	0.6172	**0.6207**
	Q_P	0.4471	0.4166	0.4712	0.4780	0.4996	**0.5132**
	Q_W	0.7051	0.6649	0.7164	0.7508	0.7679	**0.7689**
Medical	*SD*	58.3636	65.1178	64.3541	64.1187	66.6183	**68.8031**
	EN	6.0094	5.7777	5.8223	5.9089	6.0170	**6.2298**
	Q_G	0.5167	0.5992	0.5877	0.6088	0.6133	**0.6206**
	Q_P	0.3219	0.4019	0.4081	0.4139	0.4374	**0.4542**
	Q_W	0.7055	0.7504	0.7475	0.7592	0.7649	**0.7734**

Table 6
Objective assessment of the NSCT, SR and NSCT-SR methods.

Images	Metrics	NSCT	SR	NSCT-SR-1	NSCT-SR-2	NSCT-SR-3	NSCT-SR-4
Multi-focus	*SD*	52.2103	51.9298	52.0193	51.9705	52.0329	**52.2146**
	EN	7.3131	7.2997	7.3074	7.3089	7.3119	**7.3165**
	Q_G	0.7580	0.7666	**0.7702**	0.7657	0.7618	0.7582
	Q_P	0.9029	0.9064	**0.9137**	0.9088	0.9032	0.9031
	Q_W	0.9274	0.9264	**0.9294**	0.9287	0.9284	0.9274
Visible-infrared	*SD*	36.0833	40.5449	42.7409	44.1395	44.8905	**45.1291**
	EN	6.7048	6.8404	6.9579	7.0951	**7.1902**	7.1818
	Q_G	0.6510	0.6020	0.6383	0.6494	0.6581	**0.6676**
	Q_P	0.5225	0.4166	0.4962	0.5267	0.5399	**0.5482**
	Q_W	0.7343	0.6649	0.7237	0.7418	0.7629	**0.7792**
Medical	*SD*	59.9674	65.1178	65.1707	66.3673	67.6005	**68.8040**
	EN	5.7358	5.7777	5.8226	5.8408	5.8816	**5.9307**
	Q_G	0.6104	0.5992	0.6312	0.6391	0.6407	**0.6432**
	Q_P	0.4687	0.4019	0.4271	0.4414	0.4582	**0.4767**
	Q_W	0.7459	0.7504	0.7723	0.7849	0.7953	**0.7985**

MST-SR-1 methods (the MST is DWT, DTCWT, or NSCT) can generally make the best balance. On one hand, they can clearly outperform the MST methods on Q_G, Q_P and Q_W. With 1-level decomposition, the MST-SR-1 method is much more robust to mis-registration than the MST methods with 4-level decomposition. Meanwhile, the spatial details which are not fully extracted by the MST can be mostly compensated by the SR-based approach used in low-pass fusion. On the other hand, they can outperform the SR method on four (DWT-SR-1) or all the five (DTCWT-SR-1 and NSCT-SR-1) metrics. With 1-level decomposition, the MST-SR-1 methods can extract more details than the SR method. Furthermore, although the MST-SR-1 methods are more sensitive to mis-registration when compared with the SR method, there is even a slight advantage of the MST-SR-1 methods on metrics Q_G, Q_P and Q_W (This is mainly because that Q_G, Q_P and Q_W are also influenced by the spatial details preserved in the fused image. Moreover, the SR-based approach used in the low-pass fusion can also somewhat overcome the sensitivity to mis-registration).

For visible-infrared image fusion, the situation is much simpler than that of multi-focus image fusion. There are four obvious regularities we can find from Tables 1–6. First, the *SD* and *EN* of all the six MST methods are clearly lower than those of other methods, which indicates that the MST methods suffer from low contrast. Thus, the first shortcoming of the MST methods mentioned in Section 3 is verified. Second, the SR method does not work well on almost all the five metrics. Particularly, the Q_G, Q_P and Q_W of the SR method are even lower than those of some MST methods. Thus, the SR method may not be very effective for this category of image fusion for some reasons (cannot be explained by the objective performance right now). Third, the performances of the RP and RP-SR methods are poor, which means that they may also lose effectiveness in this case. Fourth, the other MST-SR methods generally exhibit a clear advantage over both the MST and SR methods on all the five metrics. Particularly, the advantage will become more obvious when the decomposition level increases.

The situation of medical image fusion is very similar to that of visible-infrared image fusion. The MST-SR methods can generally outperform both the MST and SR methods. Furthermore, from each table except for Table 2, it can be seen that the MST-SR-4 method clearly beats all the other methods in terms of all the five metrics.

To make a better comparison, we pick out eight methods from Tables 1–6 for each category of image fusion. The first two are the SR method and a specific MST method which owns the best performance among all the six MST methods. The other six are obtained by selecting the MST-SR method which has the optimal decomposition level among $\{1, 2, 3, 4\}$ from each table. Tables 7–9 list the selected results for multi-focus, visible-infrared and medical image fusion, respectively. We can see that the optimal decomposition level is 1 for multi-focus image fusion while 4 for the other two types. Furthermore, we can obtain a more meaningful outcome that the NSCT-SR-1, DTCWT-SR-4 and LP-SR-4 methods can generally achieve the best performances over

Table 7
Objective assessment of the eight selected methods for multi-focus image fusion.

Metrics	NSCT	SR	LP-SR-1	RP-SR-1	DWT-SR-1	DTCWT-SR-1	CVT-SR-1	NSCT-SR-1
SD	**52.2103**	51.9298	52.2052	52.0818	52.0612	52.0533	52.0619	52.0193
EN	7.3131	7.2997	7.3024	7.3020	7.3121	7.3071	**7.3155**	7.3074
Q_G	0.7580	0.7666	0.7650	0.7645	0.7668	0.7688	0.7630	**0.7702**
Q_P	0.9029	0.9064	0.9095	0.8962	0.9098	0.9114	0.9086	**0.9137**
Q_W	0.9274	0.9264	0.9253	0.9257	0.9259	0.9267	0.9259	**0.9294**

Table 8
Objective assessment of the eight selected methods for visible-infrared image fusion.

Metrics	NSCT	SR	LP-SR-4	RP-SR-4	DWT-SR-4	DTCWT-SR-4	CVT-SR-4	NSCT-SR-4
SD	36.0833	40.5449	45.1038	45.0224	44.1496	**45.2558**	44.6091	45.1291
EN	6.7048	6.8404	7.1237	7.1170	7.0965	7.1248	7.1132	**7.1818**
Q_G	0.6510	0.6020	0.6698	0.4782	0.6191	**0.6738**	0.6207	0.6676
Q_P	0.5225	0.4166	0.5514	0.3763	0.4547	**0.5646**	0.5132	0.5482
Q_W	0.7343	0.6649	**0.7960**	0.6066	0.7592	0.7940	0.7689	0.7792

Table 9
Objective assessment of the eight selected methods for medical image fusion.

Metrics	LP	SR	LP-SR-4	RP-SR-4	DWT-SR-4	DTCWT-SR-4	CVT-SR-4	NSCT-SR-4
SD	64.4253	65.1178	72.7315	**74.4538**	67.4361	69.8551	68.8031	68.8040
EN	5.5393	5.7777	5.8372	5.9122	5.9854	5.9832	**6.2298**	5.9307
Q_G	0.6258	0.5992	**0.6459**	0.4240	0.6193	0.6248	0.6206	0.6432
Q_P	0.4613	0.4019	**0.4905**	0.3264	0.4437	0.4583	0.4542	0.4767
Q_W	0.7678	0.7504	**0.7996**	0.5985	0.7627	0.7818	0.7734	0.7985

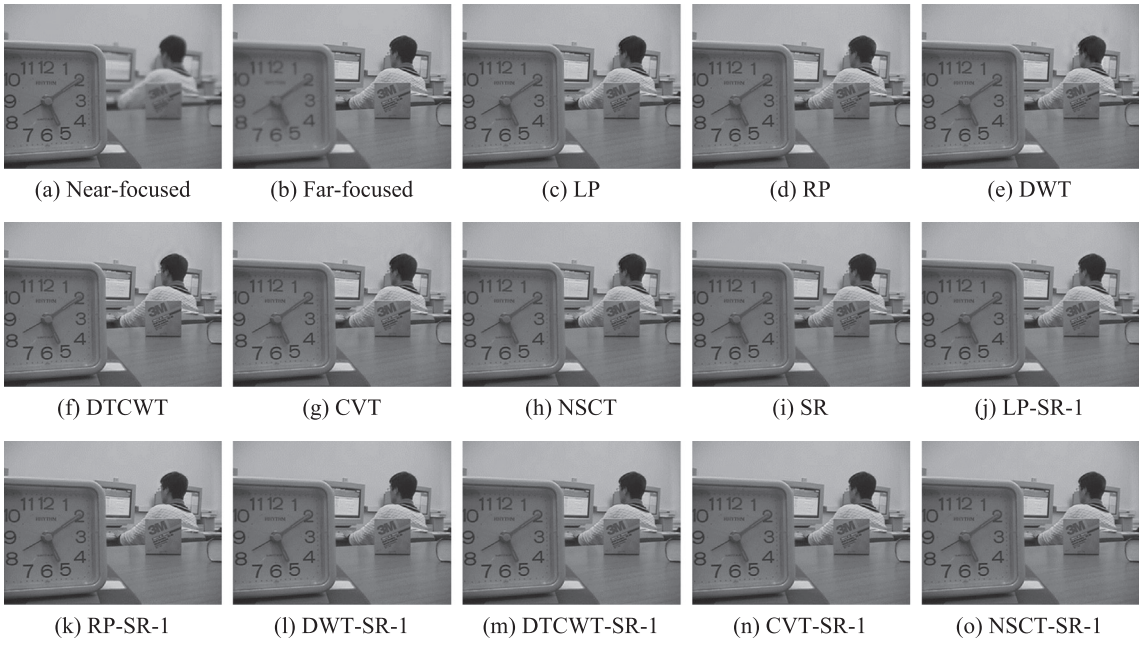


Fig. 4. The first example of multi-focus image fusion.

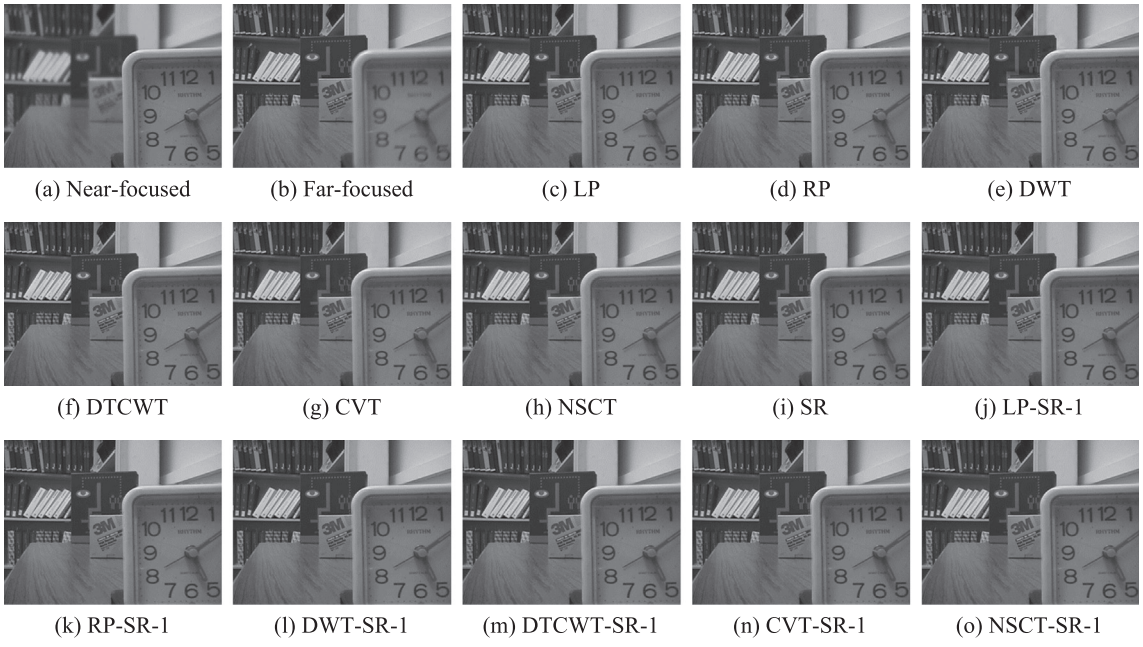


Fig. 5. The second example of multi-focus image fusion.

all the methods for multi-focus, visible-infrared and medical image fusion, respectively.

In addition to the quantitative evaluation, we also compare the fused results of different methods visually. For each category of source images, two fusion examples are given. In each example, the six MST methods, the SR method and the six MST-SR methods with the optimal decomposition level are exhibited.

Figs. 4 and 5 show two popular examples of multi-focus image fusion. In Fig. 4, there is a slight motion of the student's head in the scene. We can see that the fused images of all the six MST methods suffer from undesirable artifacts in that region. The SR and MST-SR-1 methods can obtain results in high visual quality. In Fig. 5, although there is no moving objects, the MST-SR-1 methods can obtain the fused images with more natural edges (see the white books in the bookshelf).

Two visible-infrared image fusion examples are shown in Figs. 6 and 7. It can be seen that the visible images mainly capture the bright objects, while thermal objects such as the pedestrians and plants can be easily distinguished from the infrared image. Obviously, the fused images of MST-SR-4 methods always enjoy much higher contrast than those of MST methods. The SR method is not very effective in this case. On one hand, some spatial details are lost (see the plants in Fig. 6 and the buildings in Fig. 7). On

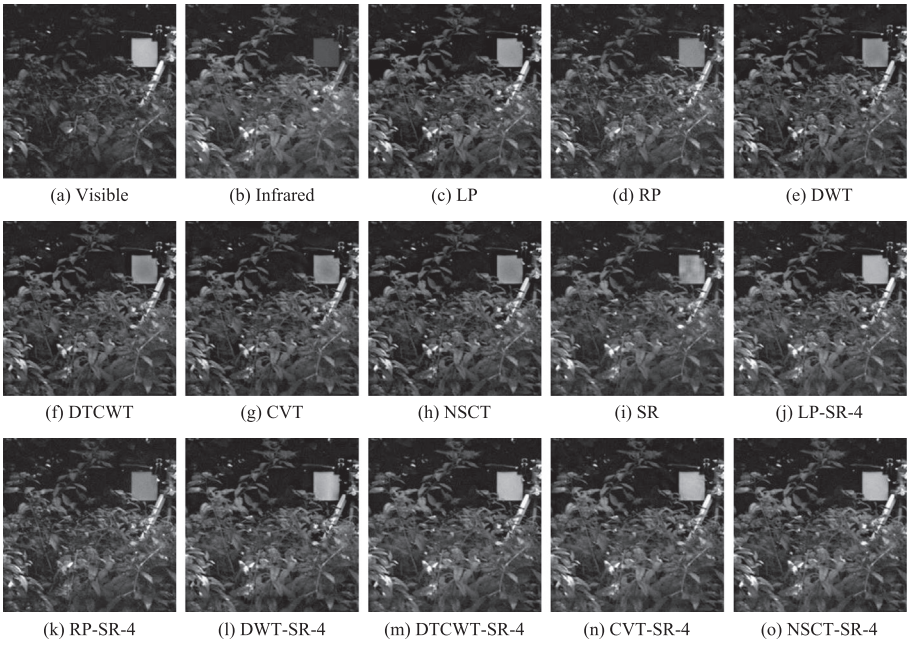


Fig. 6. The first example of visible-infrared image fusion.

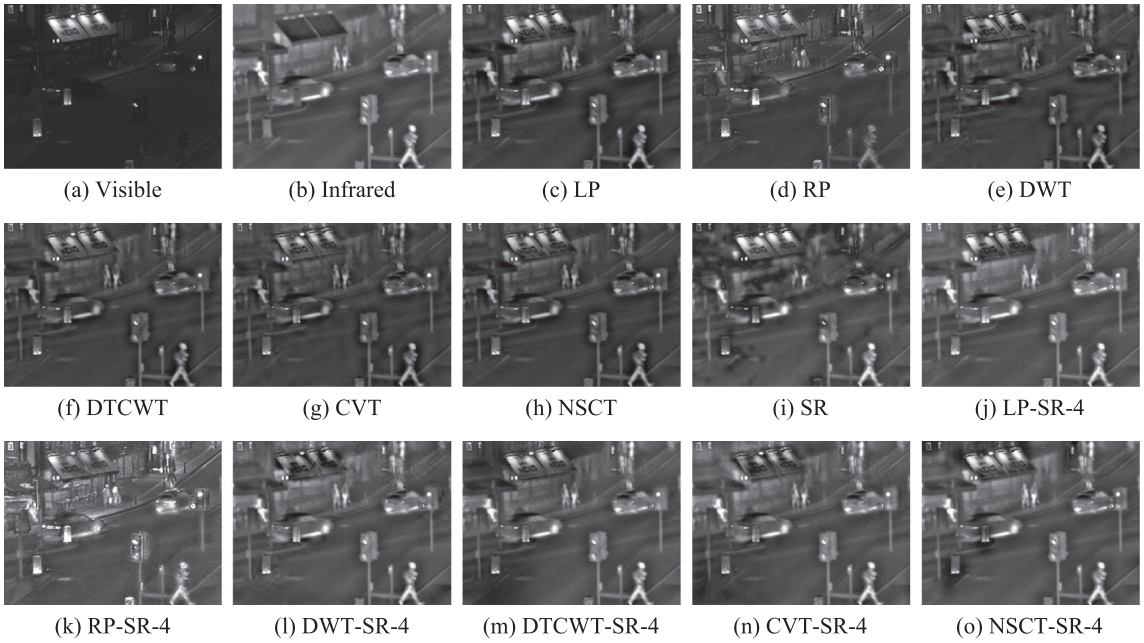


Fig. 7. The second example of visible-infrared image fusion.

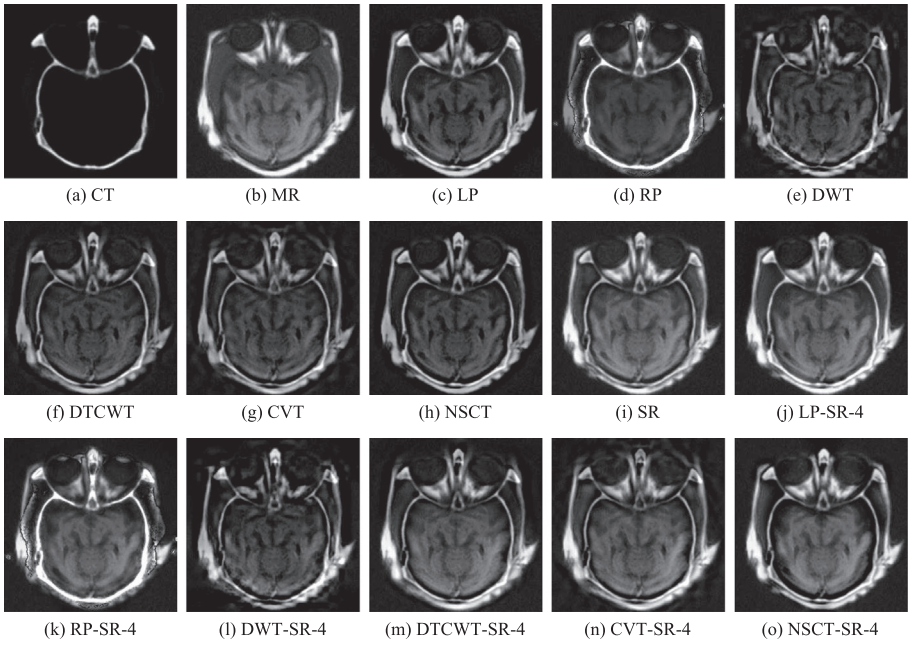


Fig. 8. The first example of medical image fusion.

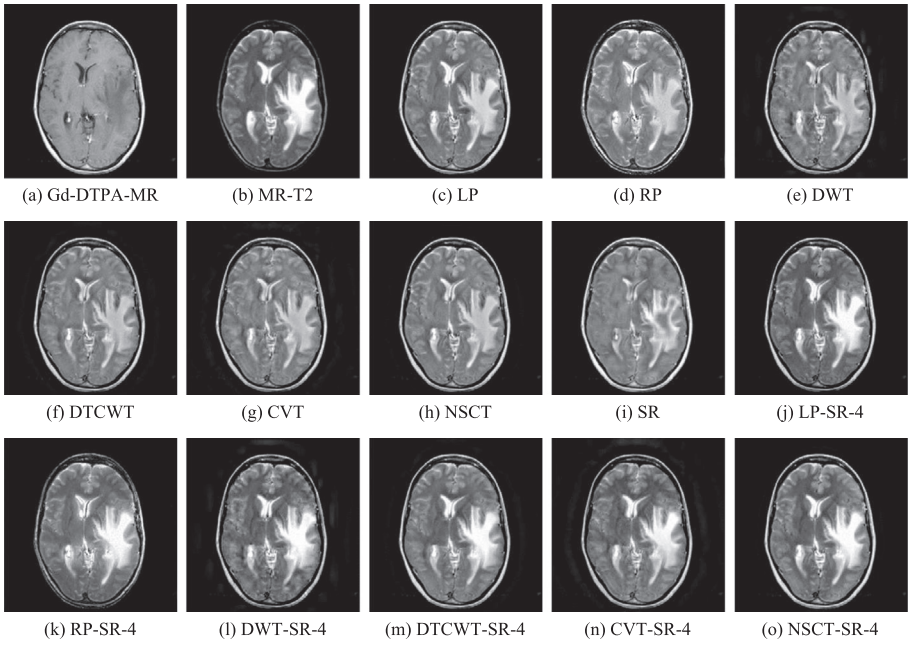


Fig. 9. The second example of medical image fusion.

Table 10
Objective assessment of the SR and NSCT-SR-1 methods with different step lengths for multi-focus image fusion.

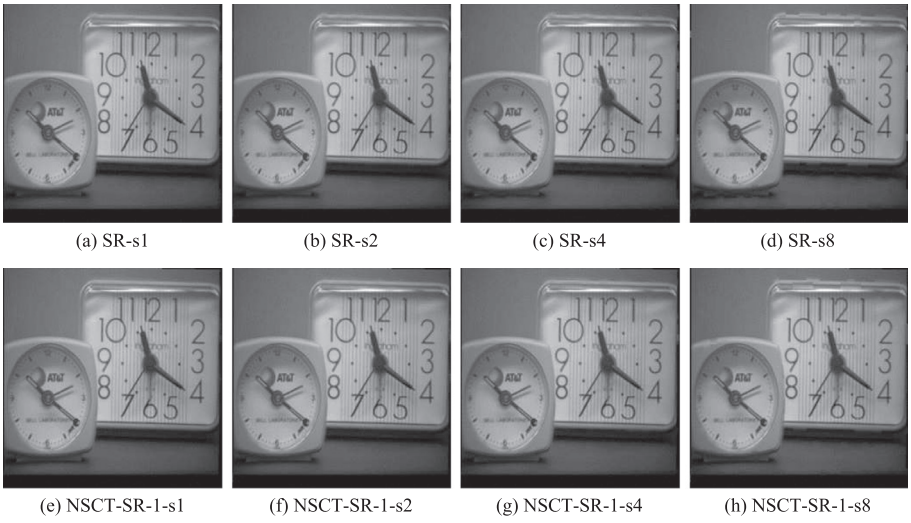
Metrics	SR-s1	SR-s2	SR-s4	SR-s8	NSCT-SR-1-s1	NSCT-SR-1-s2	NSCT-SR-1-s4	NSCT-SR-1-s8
SD	51.9298	51.9308	51.9461	52.0250	52.0193	52.0196	52.0339	**52.0635**
EN	7.2997	7.2987	7.2963	7.2785	**7.3074**	7.3067	7.3070	7.3037
Q_G	0.7666	0.7654	0.7631	0.7617	**0.7702**	0.7691	0.7669	0.7633
Q_P	0.9064	0.9021	0.8897	0.8622	**0.9137**	0.9116	0.9049	0.8841
Q_W	0.9264	0.9259	0.9247	0.9205	**0.9294**	0.9292	0.9276	0.9247
T	62,001	15,625	3969	1024	62,001	15,625	3969	1024
Time/s	375	94.5	24.1	6.10	352	89	22.9	5.99

Table 11
Objective assessment of the SR and DTCWT-SR-4 methods with different step lengths for visible-infrared image fusion.

Metrics	SR-s1	SR-s2	SR-s4	SR-s8	DTCWT-SR-4-s1	DTCWT-SR-4-s2	DTCWT-SR-4-s4	DTCWT-SR-4-s8
SD	40.5449	40.7993	42.0249	46.0220	45.2558	45.6764	46.3460	**48.0796**
EN	6.8404	6.8504	6.8757	6.8955	**7.1248**	7.1246	7.1074	7.0967
Q_G	0.6020	0.5920	0.5715	0.5269	**0.6738**	0.6717	0.6621	0.6557
Q_P	0.4166	0.3941	0.3423	0.3072	**0.5646**	0.5614	0.5568	0.5258
Q_W	0.6649	0.6609	0.6516	0.6359	**0.7940**	0.7928	0.7845	0.7664
T	62,001	15,625	3969	1024	625	169	49	16
Time/s	412	102	26.2	6.80	4.77	1.49	0.62	0.38

Table 12
Objective assessment of the SR and LP-SR-4 methods with different step lengths for medical image fusion.

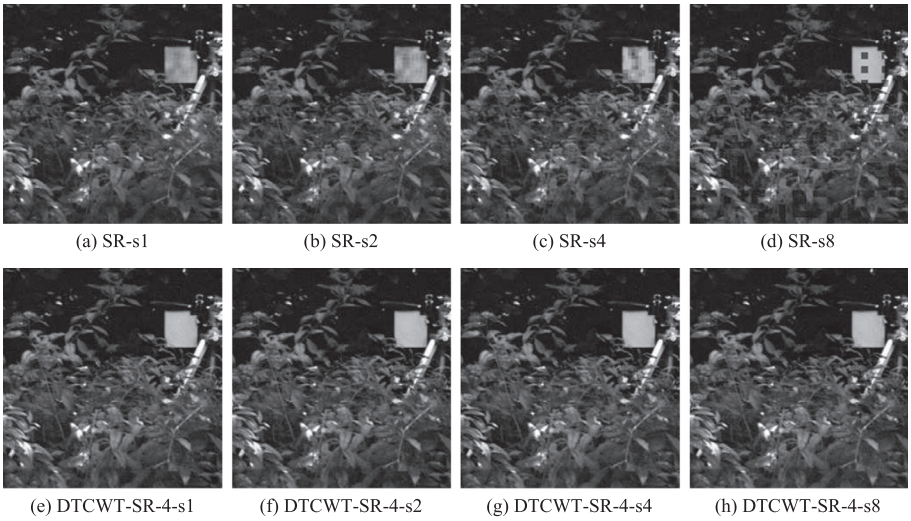
Metrics	SR-s1	SR-s2	SR-s4	SR-s8	LP-SR-4-s1	LP-SR-4-s2	LP-SR-4-s4	LP-SR-4-s8
SD	65.1178	65.3619	66.1954	69.6122	72.7315	72.7981	**73.8162**	73.6082
EN	5.7777	5.7527	5.7052	5.5537	**5.8372**	5.7460	5.7059	5.6860
Q_G	0.5992	0.5876	0.5633	0.5259	**0.6459**	0.6447	0.6436	0.6375
Q_P	0.4019	0.3728	0.3009	0.2719	**0.4905**	0.4888	0.4869	0.4829
Q_W	0.7504	0.7490	0.7399	0.7110	**0.7996**	0.7977	0.7969	0.7894
T	62,001	15,625	3969	1024	81	25	9	4
Time/s	275	69.2	17.6	4.50	0.62	0.21	0.10	0.06



(a) SR-s1 (b) SR-s2 (c) SR-s4 (d) SR-s8

(e) NSCT-SR-1-s1 (f) NSCT-SR-1-s2 (g) NSCT-SR-1-s4 (h) NSCT-SR-1-s8

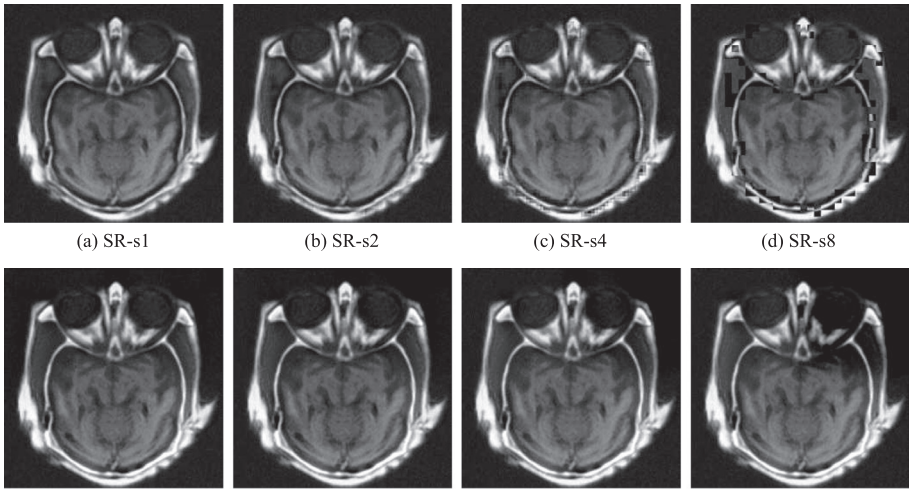
Fig. 10. A multi-focus image fusion example with different step lengths.



(a) SR-s1 (b) SR-s2 (c) SR-s4 (d) SR-s8

(e) DTCWT-SR-4-s1 (f) DTCWT-SR-4-s2 (g) DTCWT-SR-4-s4 (h) DTCWT-SR-4-s8

Fig. 11. A visible-infrared image fusion example with different step lengths.



(a) SR-s1 (b) SR-s2 (c) SR-s4 (d) SR-s8

(e) LP-SR-4-s1 (f) LP-SR-4-s2 (g) LP-SR-4-s4 (h) LP-SR-4-s8

Fig. 12. A medical image fusion example with different step lengths.

Source image 1 Source image 2 GF NSCT-SR-1

Fig. 13. The fused results of the GF and NSCT-SR-1 methods on multi-focus images.

the other hand, the spatial consistency of some regions in the fused images is destroyed (see the square board in Fig. 6 and the street in Fig. 7). Except for the RP-SR-4 method, other MST-SR-4 methods can well preserve the spatial consistency.

Figs. 8 and 9 show two fusion examples of multimodal medical images. The CT and MR images shown in Fig. 8(a) and (b) have been introduced before. The two source images shown in Fig. 9(a) and (b) are MR image after Gd-DTPA and T2-weighted MR image, respectively. We can see that images captured by different modalities reflect different organ information in human body. Just as the situation of visible-infrared image fusion, the fused images of MST methods are in low contrast, while the spatial inconsistency and loss of fine details both exist in the fused images of SR method. The MST-SR-4 methods (the MST is not RP) especially for the LP-SR-4 method, can obtain satisfactory results with high contrast, sufficient fine details and good spatial consistency.

By now, we have experimentally verified most of the theoretical discussions in Section 3 via the above objective and subjective comparisons. The only aspect that has not been discussed is the computational efficiency, which will be studied in the next subsection.

4.3. Investigation of step length

In this subsection, the impact of the sliding window's step length is studied for the three best-performed fusion methods, namely, the NSCT-SR-1 method for multi-focus image fusion, the DTCWT-SR-4 method for visible-infrared image fusion, and the LP-SR-4 method for medical image fusion. The SR method is used for comparison. For both the SR and MST-SR methods, the step length is set to 1, 2, 4 and 8 pixels, respectively. For each method, a suffix '-sk' is added to its original name, where k belongs to $\{1, 2, 4, 8\}$. For example, NSCT-SR-1-s2 denotes the NSCT-SR-1 method with a step length of 2 pixels. For each category of image fusion, in addition to the average metric values of all the source images, the average running time of fusing two source images (the first pair in each category shown in Fig. 3 is employed, and the same below) of size 256×256 is also used to compare the

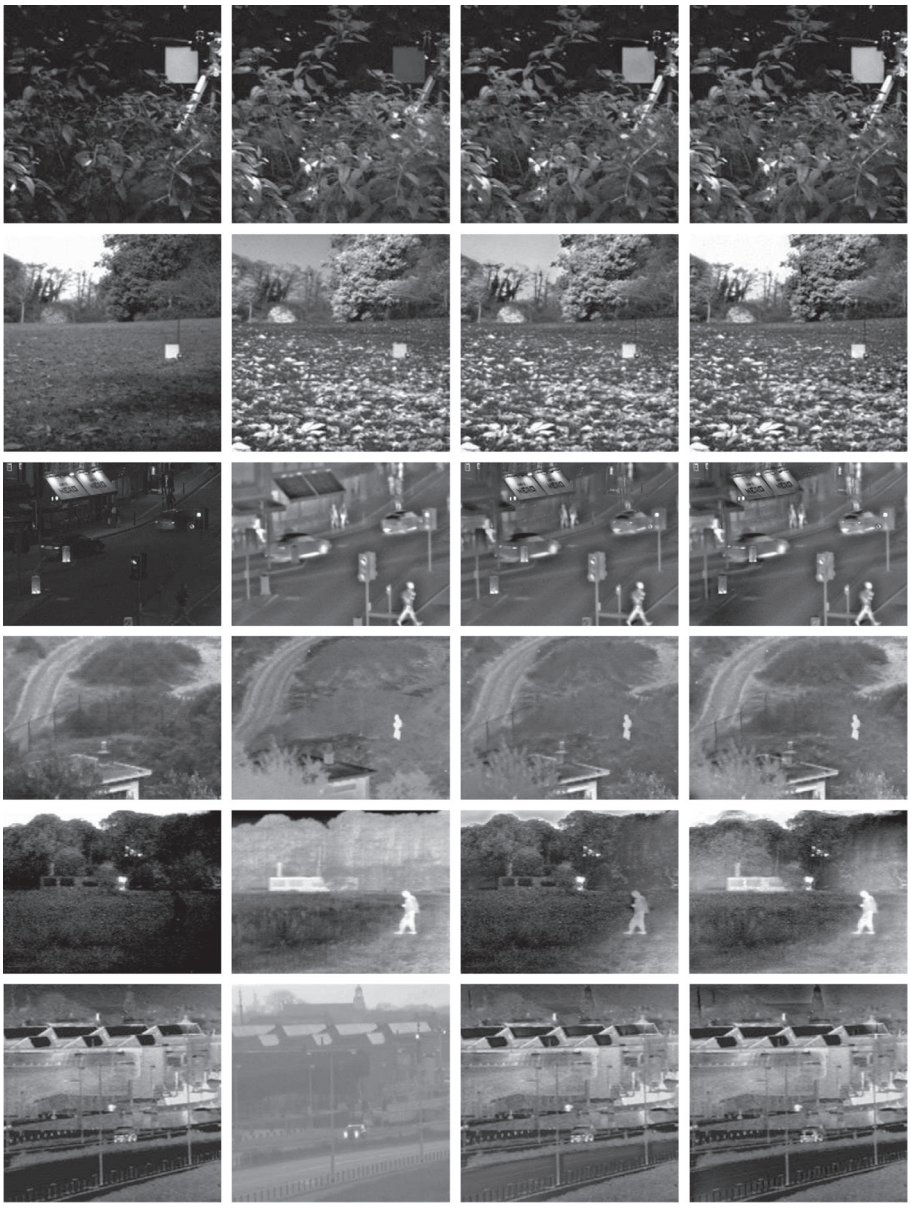


Fig. 14. The fused results of the GF and DTCWT-SR-4 methods on visible-infrared images.

performances of different fusion methods. All the fusion methods in this work are implemented in MATLAB on a computer with a 3.0 GHz CPU and 4 GB RAM.

Suppose that there is a matrix (image or low-pass band) of size $H \times W$ and the size of the sliding window is $\sqrt{n} \times \sqrt{n}$, then the number of local patches extracted from the matrix is

$$T = \left\lceil \frac{H - \sqrt{n} + 1}{s} \right\rceil \left\lceil \frac{W - \sqrt{n} + 1}{s} \right\rceil \approx \frac{(H - \sqrt{n} + 1)(W - \sqrt{n} + 1)}{s^2}, \quad (16)$$

where s is the step length of the sliding window and $\lceil \cdot \rceil$ denotes the ceiling operation. In the following three tables (Tables 10–12), we will also give the number (denoted as T) of patch pairs which are required to be fused with the sparse coding technique.

Table 10 lists the objective assessment of the SR and NSCT-SR-1 methods with different step lengths for multi-focus image fusion. The NSCT low-pass band has the same size with the original image, but we can see that the NSCT-SR-1 method is slightly efficient than the SR method with a same step length. This is mainly because the patches in the low-pass band is easier to be represented by the OMP method than those in the original image. When the step length becomes larger, all the metrics except for *SD* of both the SR and NSCT-SR-1 methods will decrease, but the computational efficiency significantly increases as expected. Furthermore, compared with the SR-s1 method, the NSCT-SR-1-s4 method have a

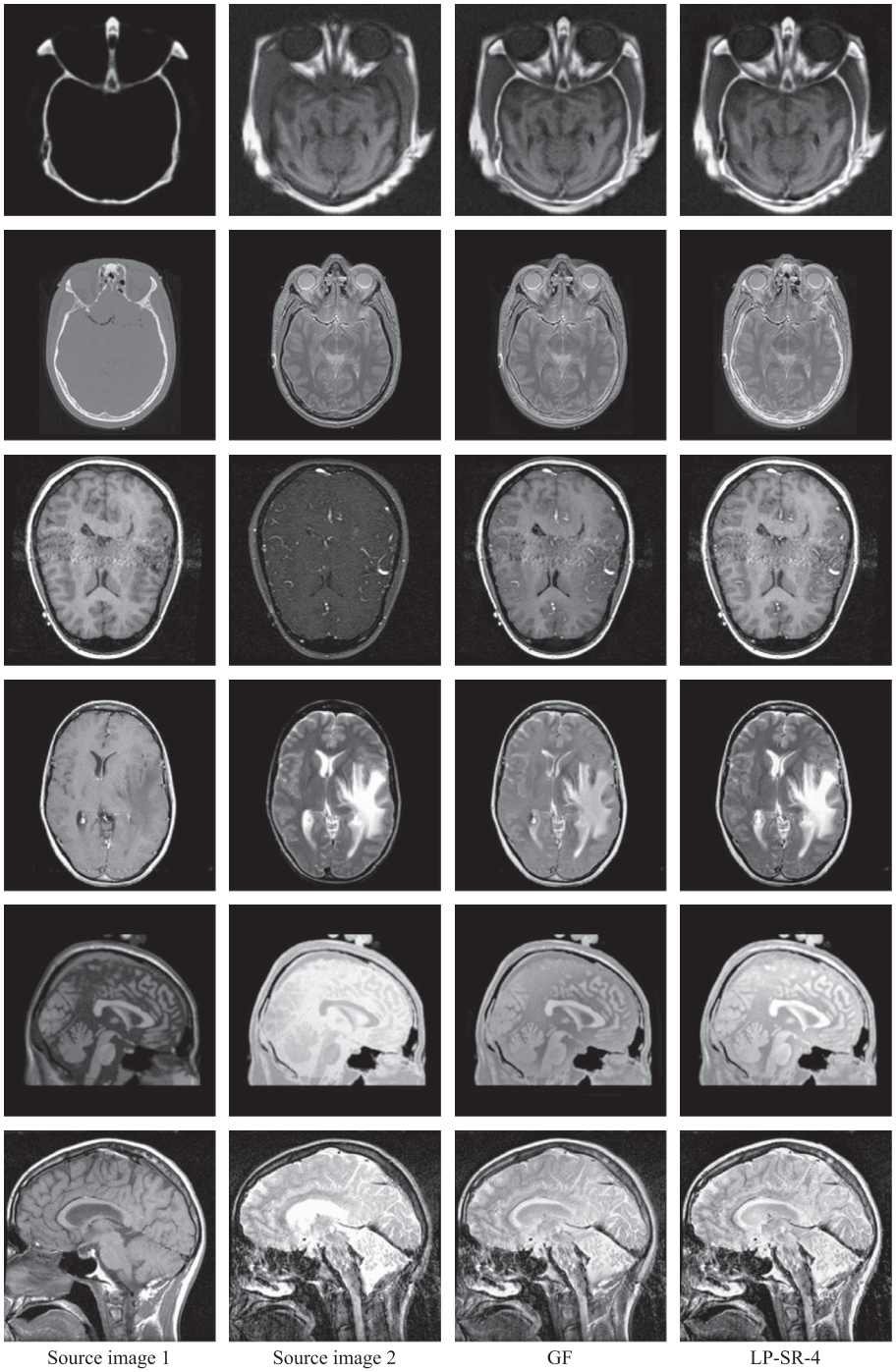


Fig. 15. The fused results of the GF and LP-SR-4 methods on medical images.

Table 13
Objective assessment of the GF and proposed fusion methods.

Images	Methods	*SD*	*EN*	Q_G	Q_P	Q_W	Time/s
Multi-focus	GF	**52.0242**	7.2965	**0.7713**	**0.9167**	0.9241	0.19
	NSCT-SR-1	52.0193	**7.3074**	0.7702	0.9137	**0.9294**	352
Visible-infrared	GF	38.4279	6.7942	0.6719	**0.5833**	0.7175	0.21
	DTCWT-SR-4	**45.2558**	**7.1248**	**0.6738**	0.5646	**0.7940**	4.77
Medical	GF	62.6755	5.8208	0.6423	**0.5296**	0.7426	0.18
	LP-SR-4	**72.7315**	**5.8372**	**0.6459**	0.4905	**0.7996**	0.62

comparative performance on Q_G, Q_P and Q_W with much less running time. A fusion example is shown in Fig. 10. It can be seen that when the step length is larger than 2, the quality of the SR method's fused results become worse. However, the fused result of the NSCT-SR-1-s4 method is still in high-quality without any obvious artifacts.

The objective assessment of the SR and DTCWT-SR-4 methods with different step lengths for visible-infrared image fusion is given in Table 11. The trends of all the five fusion metrics are same as those in Table 10. With 4-level decomposition, the size of the DTCWT low pass band is 32×32. Thus, the number of patch pairs decreases a lot so that there are only 625 patches in the low-pass band of each source image. Fig. 11 shows an example of this type of fusion. We can see that the fused results of the SR method suffer from serious blocking effect when the step length is larger than 2, while the performance of the DTCWT-SR-4 method is not very sensitive to the step length in this fusion example.

For medical image fusion, Table 12 gives the objective assessment of the SR and LP-SR-4 methods with different step lengths. The situation is very similar to that in Table 11. Since the LP low-pass band obtained with 4-level decomposition is just of size 16×16, there is a further improvement on the computational efficiency. Particularly, the LP-SR-4-s1 method only takes about 0.6 s to accomplish the fusion task. When the step length becomes larger, the running time is even shorter while the quality of the fused results is still promising. As an example shown in Fig. 12, the blocking effect in the fused images of the SR method significantly deteriorates their visual quality when the step length is 4 or 8, while the LP-SR-4-s4 method can still preserve all the important information.

Based on the above results, it can be seen that the three best-performed methods under the proposed fusion framework can still perform well so long as the step length is no more than 4. Additionally, considering that the low-pass bands of most MSTs have smaller size than the original image, the MST-SR methods are usually much more efficient than the SR method, especially when the decomposition level is 4.

4.4. Comparison with state of the art

To further evaluate the usefulness of the proposed framework, we compare the performances of the above three best-performed methods with the start of the art. In [29], Li et al. proposed an image fusion method based on guided filtering (GF). Their experimental results show that the GF method can outperform many classic and latest fusion methods for the fusion of multi-focus and multimodal images, leading to state-of-the-art results. Furthermore, the GF method has a very high computational efficiency. Thus, we apply the GF method to make a comparison on all the source images in Fig. 3. The code of the GF method is available on website [30], and all the parameters are set to the recommended values reported in [29].

For multi-focus, visible-infrared and medical image fusion, the fused results of the GF method and the corresponding proposed method are partially shown in Figs. 13–15, respectively. There are six examples for each type of fusion with each example being arranged in a row. We can see from Fig. 13 that the visual difference between the results of the two methods in multi-focus image fusion is very small. For visible-infrared image fusion, the proposed DTCWT-SR-4 method is at least comparable with the GF method. For the first or fifth example shown in Fig. 14, the DTCWT-SR-4 method outperforms the GF method. However, for the third or sixth example, the GF method performs better. For medical image fusion, it can be seen from Fig. 15 that the proposed LP-SR-4 method owns clear advantages over the GF method for all the six examples. First, the contrast in our fused images is much higher than that in the fused images of the GF method. Then, some important information is lost by the GF method (see the bone regions in the second example). Finally, the GF method tends to smooth some tiny edges (see the sixth example), while the LP-SR-4 method can well preserve them.

Table 13 lists the objective assessment of the GF and proposed fusion methods. For multi-focus image fusion, the GF method has a slight advantage over the NSCT-SR-1 method. The most distinctive difference is the computational cost, for which the NSCT-SR-1 method seems to be too inefficient. Fortunately, we can use DWT-SR-1 method or larger step length (2 or even 4) to greatly improve efficiency at a little sacrifice of the fusion quality. For the other two types of fusion, it can be seen that our methods outperform the GF method on all the five metrics except for Q_P. Although the efficiency of our methods is still lower, the gap narrows a lot, especially for the LP-SR-1 method used in medical image fusion. Moreover, as given in Table 12, the average running time of the LP-SR-1 method will decrease to 0.21 s when the step length is set to 2. Even so, the quality of the fused results can still be maintained.

5. Conclusion

In this paper, we present a general image fusion framework with multi-scale transform (MST) and sparse representation (SR). In the framework, the low-pass MST bands are merged with the SR-based scheme while the high-pass bands are fused using the conventional "max-absolute" rule. The advantages of the proposed fusion framework over conventional MST- and SR-based methods are first analyzed theoretically, and then experimentally verified. In our experiments, six popular multi-scale transforms (LP, RP, DWT, DTCWT, CVT and NSCT) with different decomposition levels ranging from one to four are first employed for the fusion of multi-focus, visible-infrared and medical images, respectively. Then, the impact of the sliding window's step length is studied. In the final, we compare our fused results with state-of-the-art level. Some main conclusions and contributions of this paper are briefly summarized as follows.

For multi-focus image fusion, the proposed MST-SR based methods can improve algorithm's robustness to mis-registration via 1-level decomposition. Meanwhile, it can overcome the inclination of the SR-based method to smooth fine details with MST. For multimodal image fusion, our fusion framework can not only obtain higher contrast than the MST methods, but also extract

more fine details and preserve better spatial consistency than the SR-based method. Furthermore, the proposed method is more efficient than the SR-based method since the time-consuming sparse coding technique is performed on low-pass bands, which may own a smaller size as well as an allowable enlargement for the step length. In particular, we give a best-performed method under the proposed framework for each category of image fusion, namely, the NSCT-SR-1 method for multi-focus image fusion, the DCTWT-SR-4 method for visible-infrared image fusion, and the LP-SR-4 method for medical image fusion. Comparisons with the latest GF-based method demonstrate that these three specific MST-SR based methods can obtain state-of-the-art results. Particularly, we believe that the LP-SR-4 method owns great potential in medical image fusion for its simple implementation, high efficiency and good performance. An image fusion toolbox which contains the MATLAB implementation of both the proposed and the compared methods is available on http://home.ustc.edu.cn/~liuyu1.

Acknowledgements

The authors first sincerely thank the editors and anonymous reviewers for their constructive comments and suggestions, which are of great value to us. The authors would also like to thank Prof. Shutao Li and Dr. Xudong Kang from Hunan University (China), Dr. Xiaobo Qu from Xiamen University (China), and Prof. Zheng Liu from Toyota Technological Institute (Japan) for generously providing some source images and codes used in the publications [11,24,27,30]. 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] A. Goshtasby, S. Nikolov, Image fusion: advances in the state of the art, Inform. Fusion 8 (2) (2007) 114–118.
[2] P. Burt, E. Adelson, The laplacian pyramid as a compact image code, IEEE Trans. Commun. 31 (4) (1983) 532–540.
[3] A. Toet, Image fusion by a ratio of low pass pyramid, Pattern Recogn. Lett. 9 (4) (1989) 245–253.
[4] V. Petrovic, C. Xydeas, Gradient-based multiresolution image fusion, IEEE Trans. Image Process. 13 (2) (2004) 228–237.
[5] H. Li, B. Manjunath, S. Mitra, Multisensor image fusion using the wavelet transform, Graph. Models Image Process. 57 (3) (1995) 235–245.
[6] M. Beaulieu, S. Foucher, L. Gagnon, Multi-spectral image resolution refinement using stationary wavelet transform, in: Proceedings of 3rd IEEE International Geoscience and Remote Sensing Symposium, 2003, pp. 4032–4034.
[7] J. Lewis, R. OCallaghan, S. Nikolov, D. Bull, N. Canagarajah, Pixel- and region-based image fusion with complex wavelets, Inform. Fusion 8 (2) (2007) 119–130.
[8] F. Nencini, A. Garzelli, S. Baronti, L. Alparone, Remote sensing image fusion using the curvelet transform, Inform. Fusion 8 (2) (2007) 143–156.
[9] Q. Zhang, B. Guo, Multifocus image fusion using the nonsubsampled contourlet transform, Signal Process. 89 (7) (2009) 1334–1346.
[10] G. Piella, A general framework for multiresolution image fusion: from pixels to regions, Inform. Fusion 4 (4) (2003) 259–280.
[11] S. Li, B. Yang, J. Hu, Performance comparison of different multi-resolution transforms for image fusion, Inform. Fusion 12 (2) (2011) 74–84.
[12] B.A. Olshausen, D.J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature 381 (6583) (1996) 607–609.
[13] M. Elad, M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries, IEEE Trans. Image Process. 15 (2) (2006) 3736–3745.
[14] B. Yang, S. Li, Multifocus image fusion and restoration with sparse representation, IEEE Trans. Instrum. Meas. 59 (4) (2010) 884–892.
[15] B. Yang, S. Li, Pixel-level image fusion with simultaneous orthogonal matching pursuit, Inform. Fusion 13 (1) (2012) 10–19.
[16] N. Yu, T. Qiu, F. Bi, A. Wang, Image features extraction and fusion based on joint sparse representation, IEEE J. Sel. Topics Signal Process. 5 (5) (2011) 1074–1082.
[17] Y. Liu, Z. Wang, Multi-focus image fusion based on sparse representation with adaptive sparse domain selection, in: Proceedings of 7th International Conference on Image and Graphics, 2013, pp. 591–596.
[18] H. Yin, S. Li, L. Fang, Simultaneous image fusion and super-resolution using sparse representation, Inform. Fusion 14 (3) (2013) 229–240.
[19] K. Engan, S.O. Aase, J.H. Husoy, Multi-frame compression: theory and design, Signal Process. 80 (10) (2000) 2121–2140.
[20] M. Aharon, M. Elad, A. Bruckstein, K-svd: an algorithm for designing overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process. 54 (11) (2006) 4311–4322.
[21] S. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries, IEEE Trans. Signal Process. 41 (12) (1993) 3397–3415.
[22] M. Elad, I. Yavneh, A plurality of sparse representations is better than the sparsest one alone, IEEE Trans. Inf. Theory 55 (10) (2009) 4701–4714.
[23] C.S. Xydeas, V.S. Petrovic, Objective image fusion performance measure, Electron. Lett. 36 (4) (2000) 308–309.
[24] J. Zhao, R. Laganiere, Z. Liu, Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement, Int. J. Innovative Comput. Inf. Control 6 (A3) (2007) 1433–1447.
[25] 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.
[26] G. Piella, H. Heijmans, A new quality metric for image fusion, in: Proceedings of 10th International Conference on Image Processing, 2003, pp. 173–176.
[27] X. Qu, 2012. <http://www.quxiaobo.org/index.html>.
[28] 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.
[29] S. Li, X. Kang, J. Hu, Image fusion with guided filtering, IEEE Trans. Image Process. 22 (7) (2013) 2864–2875.
[30] X. Kang, 2013. <http://xudongkang.weebly.com/index.html>.