Information Fusion 14 (2013) 147–162

Contents lists available at SciVerse ScienceDirect

Information Fusion

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

Image matting for fusion of multi-focus images in dynamic scenes

Shutao Li *, Xudong Kang, Jianwen Hu, Bin Yang

College of Electrical and Information Engineering, Hunan University, Changsha 410082, China

ARTICLE INFO

Article history:
Received 13 April 2011
Received in revised form 24 July 2011
Accepted 25 July 2011
Available online 31 July 2011

Keywords:
Multi-focus image fusion
Image matting
Dynamic scenes
Morphological filtering
Focus information

ABSTRACT

In this paper, we address the problem of fusing multi-focus images in dynamic scenes. The proposed approach consists of three main steps: first, the focus information of each source image obtained by morphological filtering is used to get the rough segmentation result which is one of the inputs of image matting. Then, image matting technique is applied to obtain the accurate focused region of each source image. Finally, the focused regions are combined together to construct the fused image. Through image matting, the proposed fusion algorithm combines the focus information and the correlations between nearby pixels together, and therefore tends to obtain more accurate fusion result. Experimental results demonstrate the superiority of the proposed method over traditional multi-focus image fusion methods, especially for those images in dynamic scenes.

1. Introduction

Usually, various images of the same scene can be obtained to enhance the robustness of image processing system. However, viewing and analyzing a series of images separately are not convenient and efficient. Image fusion is an effective technique to resolve this problem by combining complementary information from multiple images into a fused image, which is very useful for human or machine perception.

Recently, many image fusion algorithms have been developed to merge multi-focus images. In general, these algorithms can be classified into two groups: transform domain fusion and spatial domain fusion [1]. The most commonly used transform domain fusion methods are based on multi-scale transform. Examples of multi-scale algorithms include the discrete wavelet transform [2], gradient pyramid [3], dual tree complex wavelet transform [4], and so on. Recently, some novel transform domain analysis methods, such as curvelet transform [5], contourlet transform [6–9], log-Gabor wavelet transform [10], support value transform [11], retina-inspired model [12] and sparse representation [13] are also applied to image fusion. The existing transform domain methods consist of the following three steps in common. First, the source images are converted into a transform domain to get the corresponding transform coefficients. The transform coefficients are then merged together according to a given fusion rule. Finally, the fused image is constructed by performing the inverse transform on the fused coefficients. Different from the transform domain fusion methods, the spatial domain methods perform the fusion process in spatial domain directly. In the last few years, researchers proposed various spatial domain fusion methods which can be divided into two categories: pixel based methods [14–16] and region based methods [17–20]. Specifically, the principle of pixel or region based scheme is selecting the clearer image pixels or regions from source images to construct the fused image.

Generally, traditional multi-focus image fusion methods can generate satisfactory results for source images captured in static scenes. However, few researchers studied the problem of multi-focus image fusion in dynamic scenes with camera movement or object motion. In the dynamic scenes, the contents in the same position of multiple source images may be different. In those positions, the transform domain methods simply fuse these coefficients which represent more salient features to generate the fused image without taking into consideration that the features are probably derived from different contents. Thus, these methods usually suffer from artifacts in the fused image because of the inconsistence of image contents. For the majority of spatial domain methods, the focus information estimated by image variance, image gradient or spatial frequency is used to determine the focused pixel or region. However, in dynamic scenes, by using focus information alone, we cannot accurately decide whether a pixel or region is blurred or not. The reason is that the pixel or region in the same location of different source images may be comprised of different contents due to camera movement or object motion. Besides, traditional pixel based and region based methods cannot obtain very accurate fusion results when the patterns in the source images become complex.

In this paper, we propose a novel multi-focus image fusion algorithm with image matting which consists of the following three

* Corresponding author. Tel.: +86 731 88822924.
E-mail addresses: shutao_li@yahoo.com.cn (Shutao Li), xudong_kang@yahoo.com (X. Kang), hujianwen1@163.com (J. Hu), yangbin01420@163.com (B. Yang).

1566-2535/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.inffus.2011.07.001

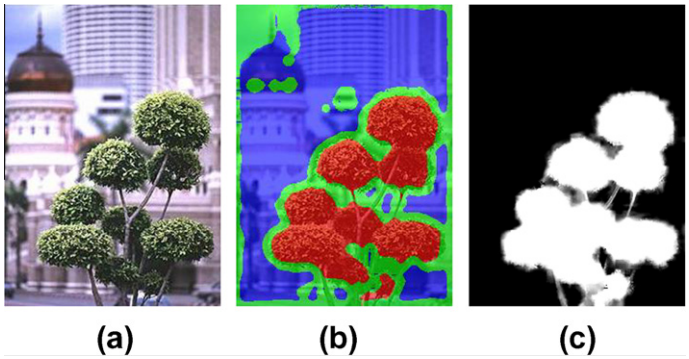

Fig. 1. The near focused "tree" source image (a), the trimap (b) and the alpha matte (c).

steps: first, morphological filtering is performed on each source image to measure focus. Then, the focus information is forwarded to image matting to find the focused object accurately. At last, the obtained focused objects of different source images are fused together to construct the fused image. The most important contribution of this paper is that the strong correlations between nearby pixels and the focus information of multi-focus images are combined together through image matting. Therefore, our algorithm can well resolve the problem of fusion of multi-focus images in dynamic scenes. Besides, another limitation of traditional spatial domain fusion methods is that their performances may degrade when image patterns become complex. On the contrary, since image matting is able to find very accurate outline of the focused object, our method can obtain very accurate fusion results in such situation. Experiments on various dynamic and static multi-focus image sets demonstrate that our method produces the state-of-the-art performance in generating satisfactory fused images, while traditional methods bring in different levels and types of undesirable artifacts.

The rest of this paper is organized as follows. In Section 2, the image matting theory is briefly reviewed. Section 3 describes the proposed image fusion algorithm in detail. The experiment results and discussions are presented in Section 4. Finally, Section 5 concludes the paper.

2. Image matting

Image matting is an important technique to accurately distinguish the foreground from background [21] which has been widely used in many applications, e.g. to obtain accurate focused object in video applications [22,23]. On account of this advantage, image matting is also adopted in the proposed method. In the model of image matting, the observed image $I(x, y)$ can be viewed as a combination of foreground $F(x, y)$ and background $B(x, y)$:

$$I(x,y) = \alpha(x,y)F(x,y) + (1 - \alpha(x,y))B(x,y), \tag{1}$$

where $\alpha(x,y)$ ranging from 0 to 1 is the foreground's opacities named as the alpha matte. $\alpha(x,y) = 1$ or 0 means that $I(x, y)$ is in foreground or background, respectively. While $\alpha(x,y)$ is a fractional value between $(0, 1)$, it means that these pixels are mixed by foreground and background. The objective of image matting is to find the accurate alpha matte so that the foreground can be accurately distinguished from the background. Since there are three unknowns F, B, α in Eq. (1), calculating the alpha matte α from a single image I is under constrained. Therefore, in most cases, the user is required to supply a trimap as the other input in addition to the original image. Fig. 1a and b shows one example image and its corresponding trimap which segments the source image into three regions: the definite foreground, the definite background and the unknown region.

In this paper, given the trimap, the resulting alpha matte is obtained by using the robust image matting algorithm [24] which consists of the following steps. First, for pixels from the unknown region shown as green in Fig. 1b, the algorithm picks out "good" samples from a set of foreground and background samples from the nearby boundaries of the foreground regions shown here as red and the nearby background regions shown here as blue.[1] In this sampling step, the confidences of every pair of foreground and background samples are estimated. In the meantime, the initial alpha matte is calculated based on Eq. (1), by selecting those pairs with high confidence as the foreground F and background B. At last, the initial estimate is further improved based on the assumption that the matte should be locally smooth and alpha values of one or zero need to be much more than mixed pixels. Mathematically, the above matting processes can be solved by an energy function which is defined as:

$$E = \sum_{z \in I} \left[\hat{f}_z(\alpha_z - \hat{\alpha}_z)^2 + (1 - \hat{f}_z)(\alpha_z - \delta(\hat{\alpha}_z > 0.5))^2 \right] + \lambda \cdot J(\alpha, a, b), \tag{2}$$

where $\hat{\alpha}_z$ and $\hat{f}_z$ are the estimated alpha and confidence values in the sampling step, δ is Boolean function returning 0 or 1, and $J(\alpha, a, b)$ is the neighborhood energy for further improving. In [24], minimizing the energy function is interpreted as solving a corresponding graph labeling problem which can be solved as a Random Walk [25]. As shown in Fig. 1c, with the given trimap, the resulting alpha matte obtained by the image matting algorithm above can accurately discriminate the focused "tree" from the source image.

3. Multi-focus image fusion based on image matting

Fig. 2 shows the schematic diagram of the proposed algorithm. First, the focus information of each source image is estimated by morphological filtering. Then, the focus information of source images are combined together to generate a trimap, which divides the corresponding source image into three regions that are the definite focused region, the definite defocused region and the unknown region. Next, with the trimap, the accurate focused region is obtained by performing image matting on the corresponding source image. Finally, the fused image is constructed by combining the focused regions of source images together.

[1] For interpretation of color in Figs. 1–3, 6, 8–11, 15, 16, the reader is referred to the web version of this article.

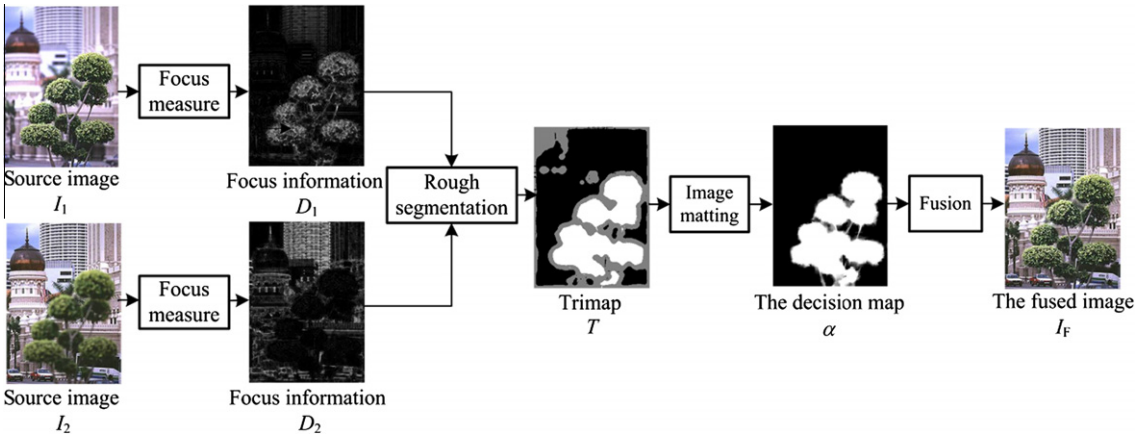

Fig. 2. Schematic diagram of the proposed algorithm.

3.1. Focus measure

The first step of our algorithm is to obtain the focus information map of each source image. It is based on the perspective that a blurred image looks "unnatural" because of the degradation of high frequency information. So, in ideal case, high frequency information is more prominent in focused region than that in defocused region. In this paper, we use morphological filtering to measure the high frequency information of source images [26]. More specifically, the morphological filtering used in this paper consists of two types of top-hat transforms

$$d_b^n = I_n - I_n \circ B, \tag{3}$$

$$d_d^n = I_n \bullet B - I_n, \tag{4}$$

where I_n is the nth source image which is obtained by weighted sum of the red, green, and blue channel, $I = 0.299 * r + 0.587 * g + 0.114 * b$ [27]. B is a disk structure element. The opening and closing operation (see Eqs. (3) and (4)) can smooth the original image by removing image's bright and dark details that do not fit within the structure element B. So, the results of top-hat transforms for opening and closing operation present the bright and dark details around the pixel (x, y). For simplicity, the maximum value of the two transforms is defined as the focus value of the corresponding pixel

$$D_n(x,y) = \max\{d_b^n(x,y), d_d^n(x,y)\}. \tag{5}$$

Fig. 3 shows one pair of source images and their corresponding focus information maps.

3.2. Rough segmentation

With the focus information obtained above, the next stage of the proposed method is to construct the trimap which roughly segments the source image into three regions, i.e., the definite focused region, the definite defocused region and the unknown region. Fig. 4 shows the schematic diagram of the rough segmentation process.

Generally, pixels from focused region more likely appear higher focus values than pixels in defocused region. In order to determine whether the pixel is blurred or not, as shown in Eq. (6), we compare the focus value of each pixel at first.

$$R_n(x,y) = \begin{cases} 1, & D_n(x,y) > \max\limits_{m,m\neq n}\{D_m(x,y)\} \\ 0 & \text{otherwise} \end{cases}. \tag{6}$$

For the source images in Fig. 3a and b, the map of focused pixels R_1 and R_2 are shown in Fig. 5a and b respectively. $R_n(x, y) = 1$ means the point (x, y) of the nth source image is in focus. However, in practice, comparing the focus value is not exactly right for these pixels in areas with no discernable edges or details and motion areas. So, there exist a proportion of pixels with the highest magnitude from out-of-focus region, and vice verse. In the stage of rough segmentation, instead of using focus information to determine the accurate focused region, we only need to find the coarse but highly believable focused region named as the definite focused region. In this paper, the obtained map of focused pixels R_n of source images are further processed by the following steps. First, a median filter is employed to remove these isolated pixels and very small regions caused by image noise, etc.

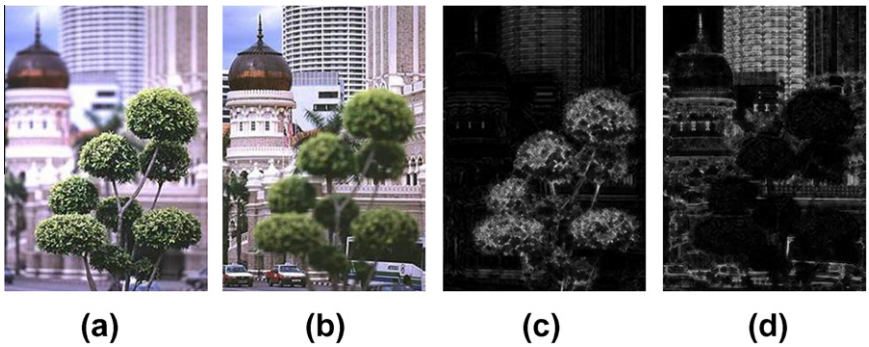

Fig. 3. The "tree" source images (a and b), and the focus information map obtained by morphological filtering (c and d).

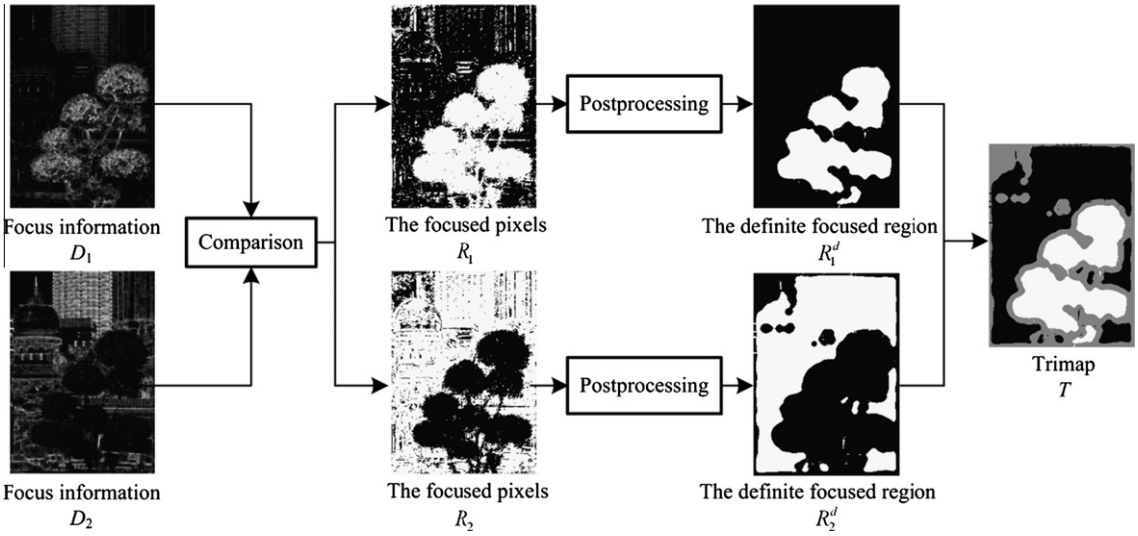

Fig. 4. Schematic diagram of the rough segmentation process.

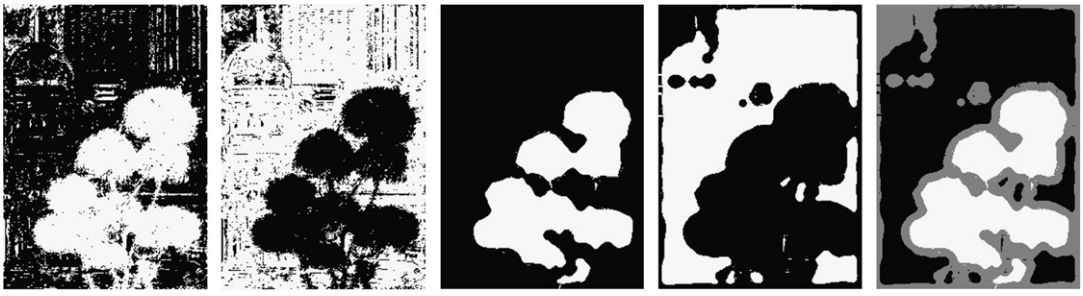

Fig. 5. The focused pixels of source images (a and b), the definite focused regions of source images (c and d) and the obtained trimap of source image I_1 (e).

$$R_n^M(x,y) = \mathrm{Med}\{R_n(x-k,y-l)|(k,l) \in w\} \tag{7}$$

where Med stands for the median filtering operation, w is a sliding local window. Then, based on the assumption that motion regions always appear in the boundary of objects caused by camera movement or in the form of scattered pieces caused by object motion, through performing several iterations of skeletonization [27] and one time of median filtering (see Eq. (8)), these scattered pieces and outer areas of the focused regions in R_n^M can be effectively removed.

$$R_n^{MS} = \mathrm{Med}\{\mathrm{Skelet}(R_n^M, i)\}, \tag{8}$$

where Skelet stands for the skeletonization operation and i denotes the number of skeletonization iterations. Besides, these pixels with very large focus value which are more likely to be focused are defined as the definite focused pixels.

$$P_n^d(x,y) = \begin{cases} 1 & \text{if } D_n(x,y) - \max\limits_{m,m\neq n}\{D_m(x,y)\} > H \\ 0 & \text{otherwise} \end{cases}, \tag{9}$$

where D denotes the focus information obtained by focus measure. H is a threshold parameter which ranges from 0 to 255 for 8-bit images. At last, the definite focused region of the nth source image is defined as

$$R_n^d(x,y) = \begin{cases} 1 & R_n^{MS}(x,y) = 1 \text{ or } P_n^d(x,y) = 1 \\ 0 & \text{otherwise} \end{cases}. \tag{10}$$

Fig. 5c and d shows the definite focused region obtained as above. Given the definite focused region of each source image, the trimap of source image I_n is defined as

$$T_n(x,y) = \begin{cases} 1, & \text{if } R_n^d(x,y) = 1 \text{ and } \max\limits_{m,m\neq n}\{R_m^d(x,y)\} = 0 \\ 0, & \text{if } R_n^d(x,y) = 0 \text{ and } \max\limits_{m,m\neq n}\{R_m^d(x,y)\} = 1 \\ 0.5 & \text{otherwise} \end{cases}. \tag{11}$$

The trimap T_1 of source image I_1 is shown in Fig. 5e, where $T_n = 1$ or $T_n = 0$ means that the point (x, y) of the nth source image is definitely in focus or not in focus, respectively. $T_n = 0.5$ means that these pixels are from the unknown region which need to be classified by image matting.

3.3. Image matting and fusion

The last stage of our proposed method is to construct the fused image by combining the focused regions of source images together. In order to obtain the focused region of each source image, taking the trimap T_n obtained above as one input, the robust image matting algorithm [24] is performed on source image I_n to get the alpha matte $\alpha_n(x,y)$. More specifically, for each pixel in the unknown region, the matting algorithm first picks out several pairs of focused and defocused samples which located along the boundaries of the definite focused region and the definite defocused region respectively. Based on the color similarity between the

Fig. 6. The alpha matte obtained by image matting (a), the focused region of each source image (b and c) and the fused image obtained by the proposed method (d).

unknown pixel and the sampled pairs of focused and defocused samples, the initial alpha value of the unknown pixel is then estimated. Finally, based on the assumption that the obtained alpha matte should be locally smooth and alpha values of one or zero (equal to focus or defocus) should to be much more common than mixed pixels (alpha is a fractional value between (0, 1)), the accurate alpha matte is calculated by minimizing a defined energy function (see Eq. (2)). Fig. 6a shows the resulting alpha matte where $\alpha_n(x,y) = 1$ or 0 means that the point (x, y) of the source image I_n is in focus or not in focus respectively. While $\alpha_n(x,y)$ is a fractional value between (0, 1), it means that these pixels are mixed by focused pixels and out-of-focus pixels. Since the number of the mixed pixels is small and they are usually located in the transition regions between focus and defocus, they will not reduce the global performance of the fused image. In this paper, the resulting alpha matte α_n is identical to the focused region of source image I_n. For two source images, the focused region of source image I_2 can be simply calculated by $1 - \alpha_1(x,y)$. Thus, the fused image of two source images can be calculated by

$$I_F(x,y) = \alpha_1(x,y)I_1(x,y) + (1 - \alpha_1(x,y))I_2(x,y). \quad (12)$$

The focused regions of source images and their corresponding fused image are presented in Fig. 6b–d, respectively. As shown in Fig. 6d, all important information has been well preserved. At last, Fig. 7 shows the algorithm description of the proposed method for better understanding and reproducibility.

Algorithm1 Multi-focus image fusion method with image matting

Step 0: Given multi-focus images I_n, $n = 1, \ldots, N$.

Step 1: Measure the focus information of each source image by morphological filtering.

Step 2: Compare the focus value of each pixels in different source images to obtain the map R_n of focused pixels of each source image I_n.

Step 3: Perform median filtering and skeletionization on the R_n obtained above to get the definite focused region R_n^d of each source image.

Step 4: Segment each source image I_n into three portions with the R_n^d obtained above. And the segmented result T_n is the trimap used for image matting.

Step 5: Perform image matting on source image I_n ($n = 1, \ldots, N$-1) to obtain the alpha matte α_n of I_n. The alpha matte indicates the focused region of I_n.

Step 6: Construct the fused image by the following fusion processes:

$$I_{n,N}(x,y) = \alpha_n(x,y)I_n(x,y) + (1 - \alpha_n(x,y))I_{n-1,N}(x,y)$$

where n refers to 1,2,…,N-1 in order. $I_{n,N}(x,y)$ is the fused image of the first n source images and the N source image. When $n = 1$, $I_{n-1,N}(x,y) = I_{0,N}(x,y) = I_N(x,y)$ which denotes the Nth source image. When $n = N - 1$, the final fusion result $I_{N-1,N}(x,y)$ is obtained.

Fig. 7. The proposed multi-focus image fusion algorithm with image matting.

4. Experiments and discussions

4.1. Experimental setup

For the proposed method, the structure element B used for two types of top-hat transforms is defined as following:

$$B = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

The median filtering is performed on an 8×8 neighborhood. When the number of pixels in the neighborhood is even, the median value is calculated by choosing the bigger one of the two medians. The iteration time i of skeletonization is set as 5. The threshold H is set as half of the magnitude of image gray level, i.e., 128 for 8-bit images. The fusion results of the proposed method are compared with six other image fusion algorithms which are based on dual tree complex wavelet (DTCWT) [4], non-subsampled contourlet transform (NSCT) [7], sparse representation (SR) [13], image gradient (IG) [17], region segmentation and spatial frequency (RGSF) [19] and mathematic morphology (MM) [26]. For the DTCWT based method, four decomposition levels, the "averaging" scheme for the low-pass sub-band and the "absolute maximum choosing" scheme for the band-pass sub-band are adopted. Four decomposition levels, together with 4, 8, 8, 16 directions from coarser scale to finer scale are used for the NSCT based method. For the sparse representation based method, the block size of sliding window is set as 8×8, and the global error is set as 0.01. The IG based method is performed by taking the parameter $k = 60$ suggested in their paper. For the RGSF based method, the most important parameter is the number n of segmented portions. We tune this parameter and choose the one which obtains the best result. For the method based on mathematic morphology, the radius of the disk structure element is also set as 2 pixels. The termination criteria $p\%$ of the algorithm is set as 95% which is suggested in their paper.

4.2. Objective evaluation

In order to assess the performance of different methods objectively, many related assessing indexes have been published and

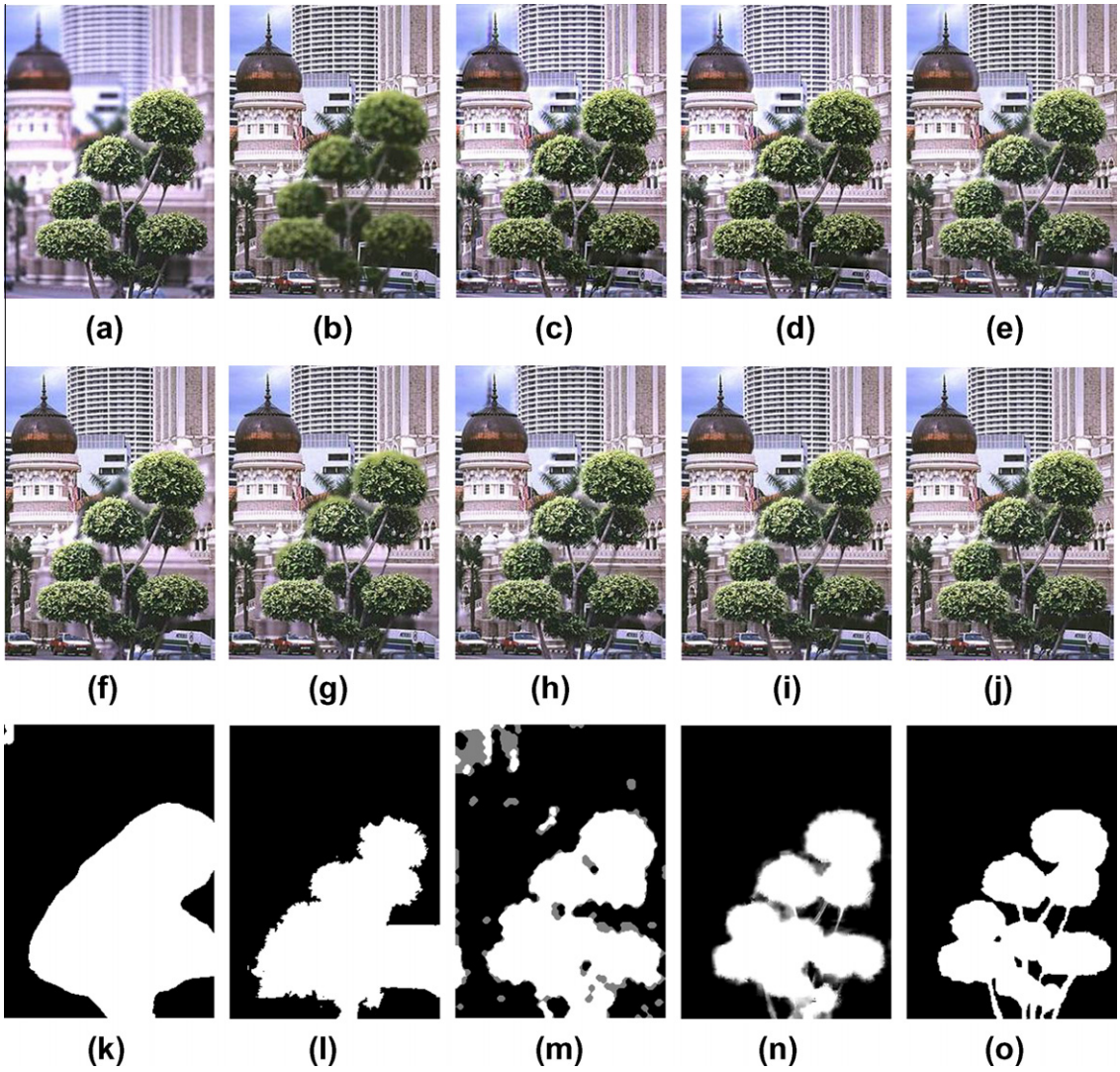

Fig. 8. The "tree" source images (a and b), the fused images by DTCWT (c), NSCT (d), SR (e), IG (f), RGSF (g), MM method (h), proposed method (i), reference fused image (j), the fusion decision maps of IG (k), RGSF (l), MM (m), proposed method (m), and reference decision map (o). (Image courtesy of Xiaoyi Jiang).

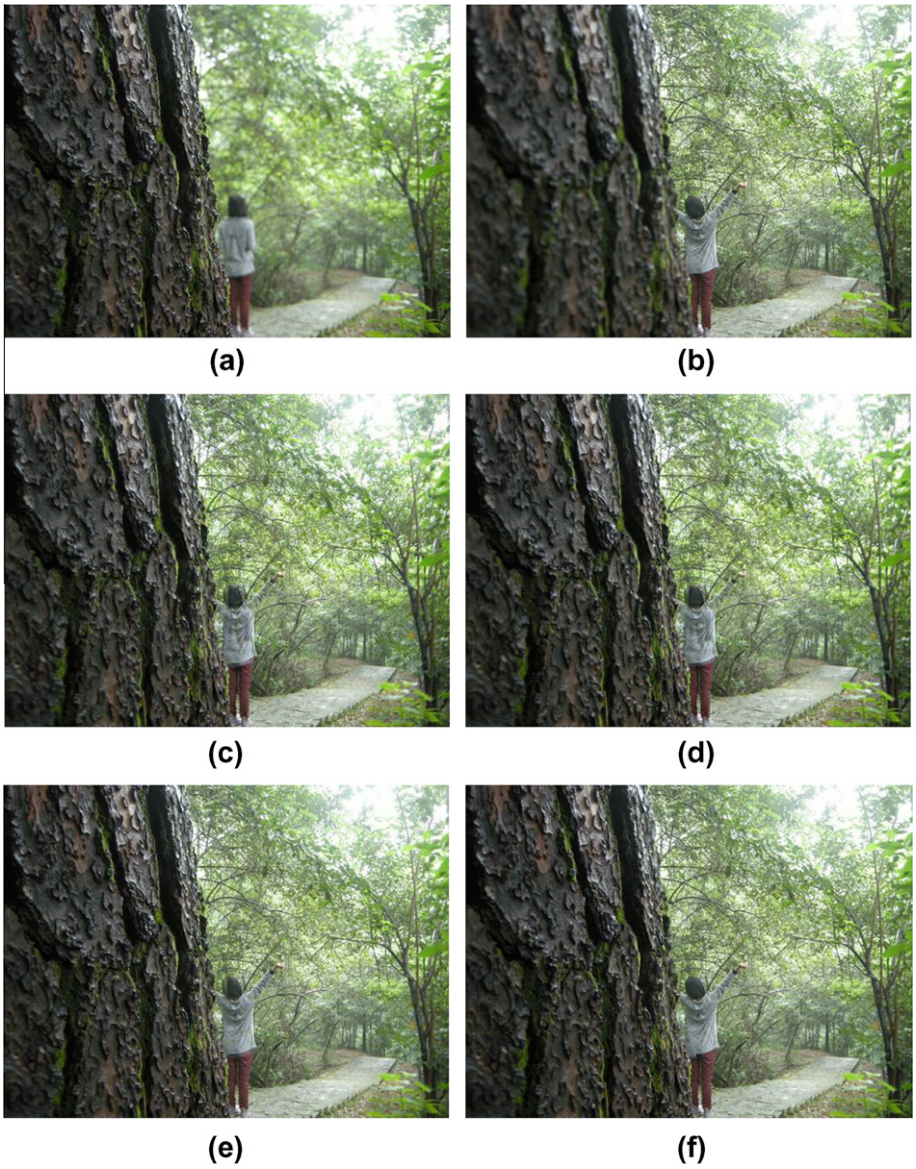

Fig. 9. The "trunk" source images (a and b), the fused images by DTCWT (c), NSCT (d), SR (e), IG (f), RGSF (g),MM (h), proposed method (i), and reference fused image (j).

discussed in references [28–31]. In this paper, the modified mutual information M_F^{XY} [32], the structural similarity metric SSIM [33], the similarity based quality metric $Q(X, Y, F)$ [34] and the gradient based metric $Q^{XY/F}$ [35] are adopted. The first index, the modified version of mutual information, is defined as:

$$M_F^{XY} = 2\left[\frac{I(F,X)}{H(F)+H(X)} + \frac{I(F,Y)}{H(F)+H(Y)}\right] \tag{13}$$

where X and Y are the two source images. F is the fused image. $I(F, X)$ and $I(F, Y)$ represent the mutual information between the fused image F and source images X, Y respectively. $H(X)$, $H(Y)$ and $H(F)$ denote the entropy of X, Y and F, respectively.

The second index is the structural similarity metric SSIM proposed by Wang et al. [33]. The mathematic expression of SSIM is:

$$\mathrm{SSIM}(R,F) = \frac{(2\mu_R\mu_F - C_1)(2\sigma_{RF} + C_2)}{(\mu_R^2 + \mu_R^2 + C_1)(\sigma_R^2 + \sigma_F^2 + C_2)}, \tag{14}$$

where R and F are the reference image and the fused image respectively. $C_1 = (K_1L)^2$, $C_2 = (K_2L)^2$, L is the dynamic range of the pixel values. K_1 and K_2 are small constants. Here, the default parameters: $K_1 = 0.01$, $K_2 = 0.03$, $L = 255$ are adopted. The reference image is created by manual segmentation and paste operation. Thus, the value of SSIM refers to the structural similarity between the fused image and the manual created reference image.

The third quality metric is the similarity based quality metric $Q(X, Y, F)$ proposed by Yang et al. [34] which performs different operations when evaluating different local regions according to the similarity level between the source images. Considering there are two source images X and Y, and one fused image F, by using a sliding window W with the size of 7×7 which moves from the top-left corner to the bottom-right corner, the local structural similarities $\mathrm{SSIM}(X, Y|W)$, $\mathrm{SSIM}(X, F|W)$ and $\mathrm{SSIM}(Y, F|W)$ are calculated respectively. Based on the measured $\mathrm{SSIM}(X, Y|W)$, the quality measure in W is defined as:

$$Q(X,Y,F|W) = \begin{cases} \lambda(W)\mathrm{SSIM}(X,F|W) + (1-\lambda(W))\mathrm{SSIM}(Y,F|W), \\ \quad \text{for } \mathrm{SSIM}(X,Y|W) \geqslant 0.75 \\ \max\{\mathrm{SSIM}(X,F|W),\ \mathrm{SSIM}(Y,F|W)\}, \\ \quad \text{for } \mathrm{SSIM}(X,Y|W) < 0.75 \end{cases} \tag{15}$$

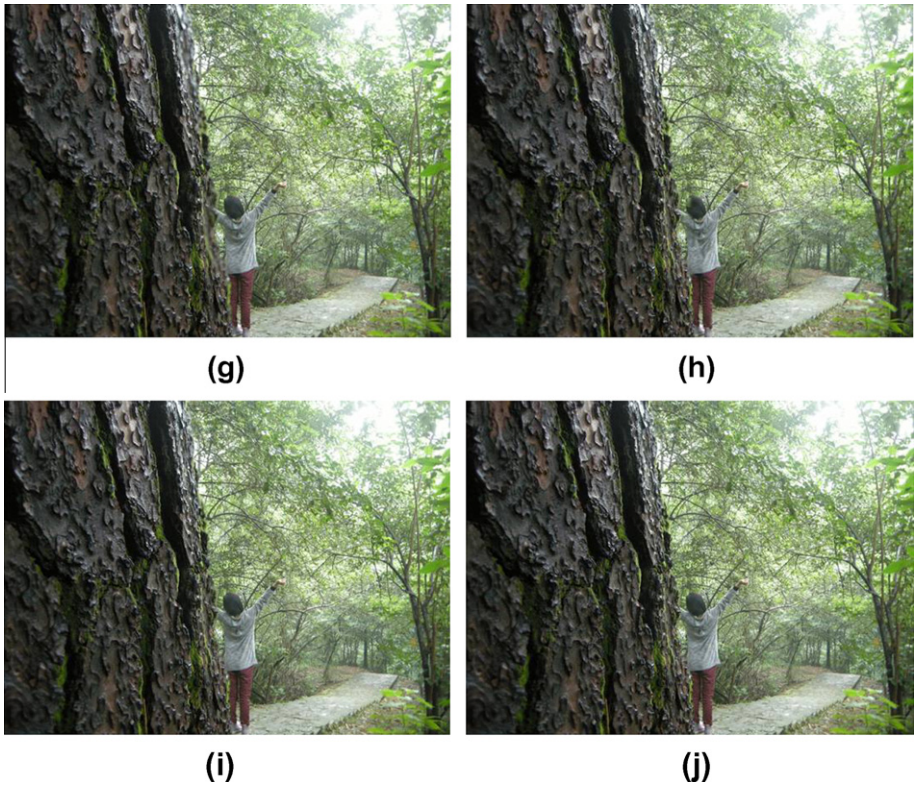

Fig. 9 (*continued*)

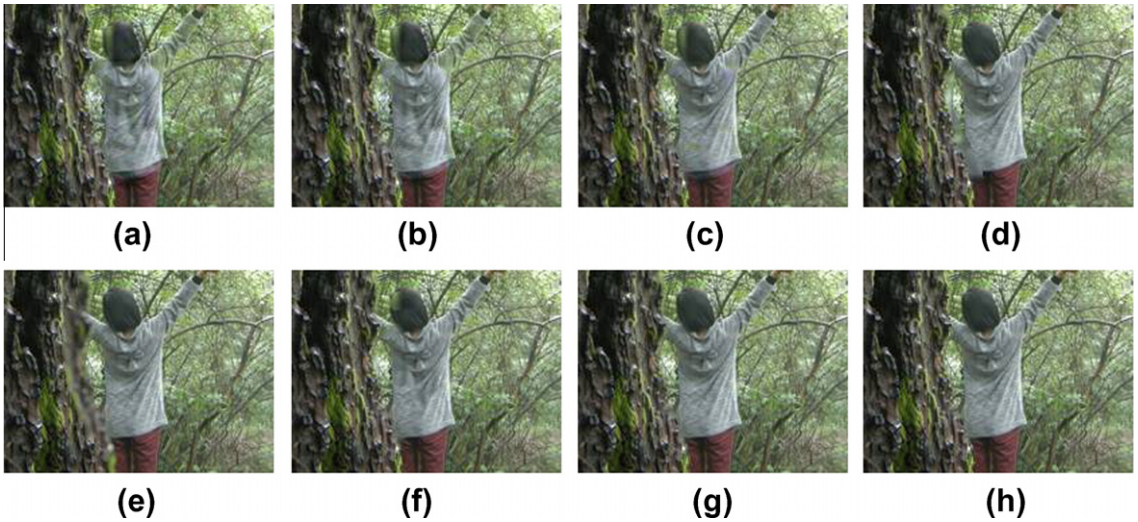

Fig. 10. Magnified regions of the fused images by DTCWT (a), NSCT (b), SR (c), IG (d), RGSF (e), MM (f), proposed method (g), and reference fused image (h).

where, $\lambda(W) = \frac{S(X|W)}{S(X|W)+S(Y|W)}$ is the local weight, and $S(X|W)$, $S(Y|W)$ are the local variances of W_X and W_Y, respectively. The global quality measure $Q(X, Y, F)$ is obtained by averaging all the values over the whole image. The parameters C_1 and C_2 used for calculating local structural similarities are set as 2^{-16} which are the default parameters given in [34].

The index $Q^{XY/F}$ proposed by Petrović and Xydeas assesses fusion performance by evaluating the amount of edge information transfer from the source images to the fused image [35]. It is defined as:

$$Q^{XY/F} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(Q^{XF}(i,j)w^X(i,j) + Q^{YF}(i,j)w^Y(i,j))}{\sum_{i}^{N}\sum_{j}^{M}(w^X(i,j) + w^Y(i,j))}, \tag{16}$$

where $Q^{XF}(i,j) = Q_g^{XF}(i,j)Q_o^{XF}(i,j)$. $Q_g^{XF}(i,j)$ and $Q_o^{XF}(i,j)$ are the Sobel edge strength and orientation preservation value at location (i,j), respectively. N and M are the size of the images. $Q^{YF}(i,j)$ is similar to $Q^{XF}(i,j)$. $w^X(i,j)$ and $w^Y(i,j)$ reflect the importance of $Q^{XF}(i,j)$ and $Q^{YF}(i,j)$, respectively. The parameters $L = 1.0$, $\Gamma_g = 0.9994$,

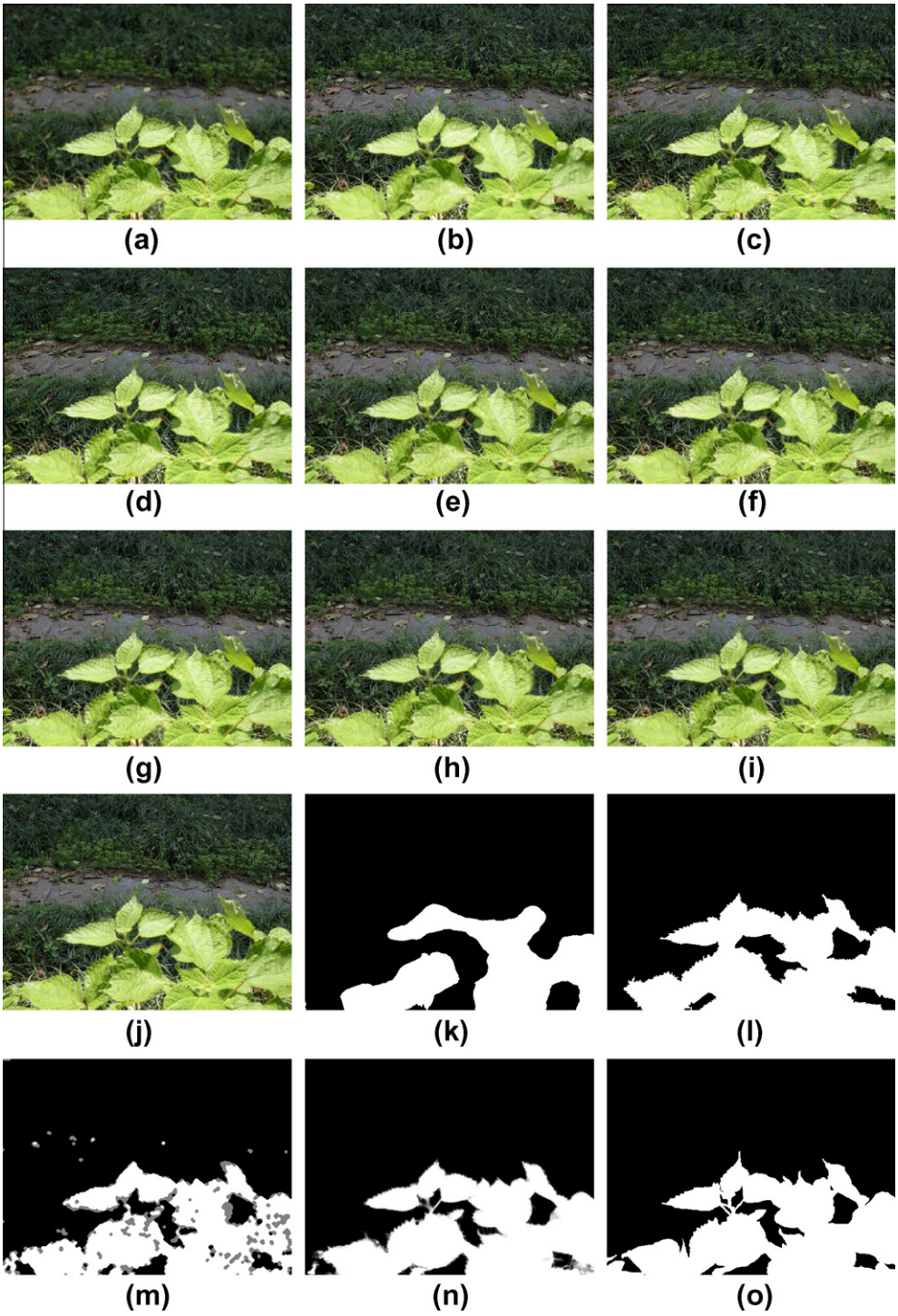

Fig. 11. The "leaves" source images (a and b), the fused images by DTCWT (c), NSCT (d), SR (e), IG (f), RGSF (g), MM method (h), proposed method (i), reference fused image (j), the fusion decision maps IG (k), RGSF (l), MM (m), proposed method (n), and reference decision map (o).

$\kappa_g = 15$, $\sigma_g = 0.5$, $\Gamma_\alpha = 0.9879$, $\kappa_\alpha = 22$ and $\sigma_\alpha = 0.8$ are adopted for this quality metric. Note that, for colored images, the average index value of three color channels is used for comparison. Besides, for all assessing indexes, the larger the index value is, the better the fusion performance is.

4.3. Fusion results

The first experiment is performed on the "tree" source images shown in Fig. 8a and b (tree and church in focus, respectively), where the building and the vehicles cannot be exactly registered due to camera movement and object motion. The fused images of different methods are presented in Fig. 8c–i. Further more, as shown in Fig. 8j, we created a reference fused image by manual segmentation. The fusion decision maps of different spatial domain methods and the reference decision map are given in Fig. 8k–o for visual comparison. The fused images of DTCWT, NSCT and SR based methods are shown in Fig. 8c–e. It can be observed that on the top of the church and on the front of the bus there are some distortions due to camera movement and motion of the vehicle. In addition,

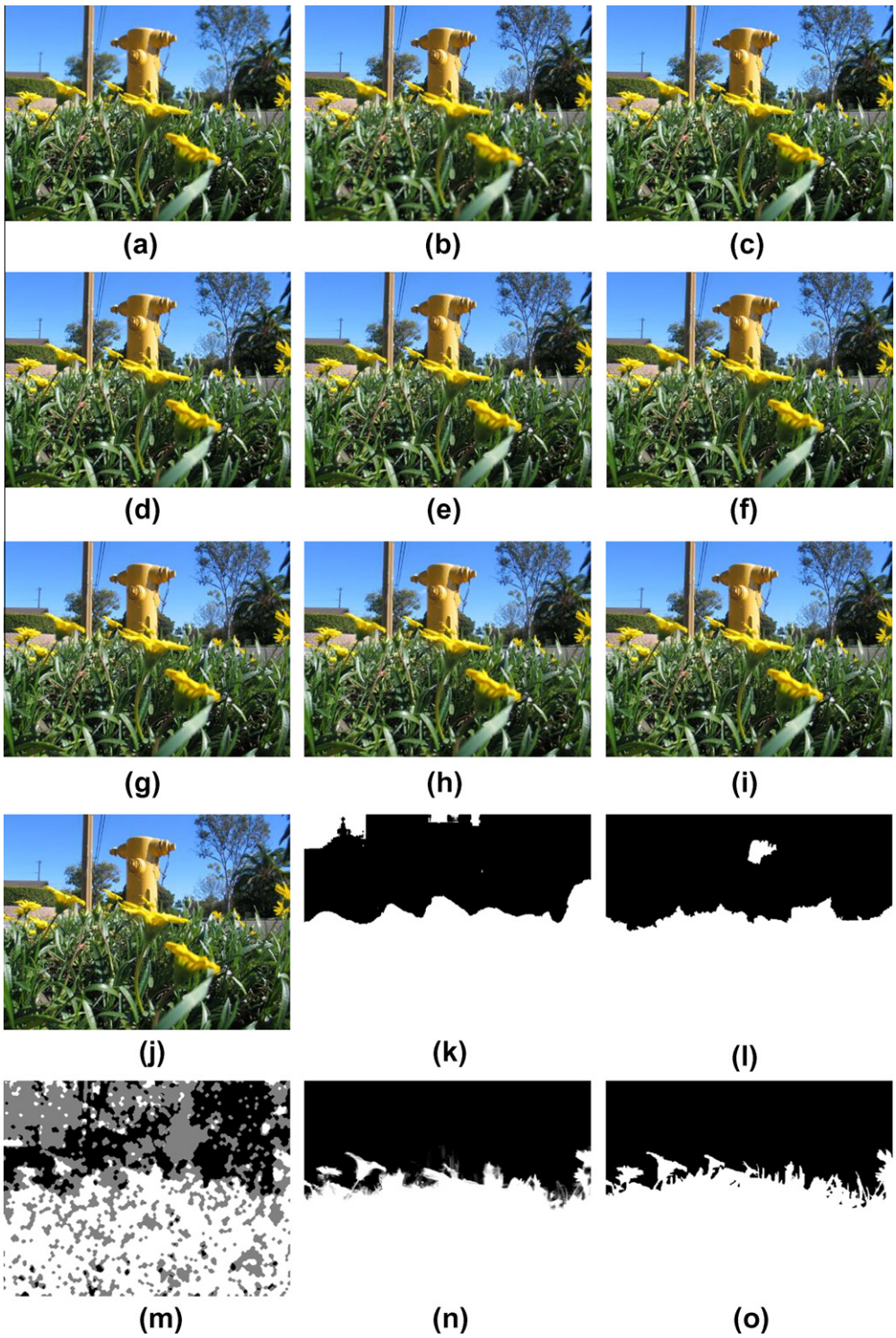

Fig. 12. The "hydrant" source images (a and b), the fused images by DTCWT (c), NSCT (d), SR (e), IG (f), RGSF (g), MM method (h), proposed method (i), reference fused image (j), the fusion decision maps IG (k), RGSF (l), MM (m), proposed method (n), and reference decision map (o) (Image courtesy of Dmitry V. Fedorov.).

the decision maps in Fig. 8k and l shows that the IG and RGSF based methods cannot obtain accurate decision maps because of the complex patterns of source images. For instance, the region among the branches is still blurred in their fused images, as shown in Fig. 8f and g. In the gray area of the decision map of the MM based method shown in Fig. 8m, the fused image is constructed by using the average value of source images. Although the region between the branches is clear in the fused image (see Fig. 8h), artifacts appear in mis-registered parts such as the top of the church and motion areas such as the front part of the bus. The fused image of the proposed method shown in Fig. 8i demonstrates that all useful information of the source images has been transferred to the fused image. Meanwhile, fewer artifacts are introduced during the fusion process. Besides, it can be observed that the decision map of our method in Fig. 8n is the closest to the reference decision map in Fig. 8o.

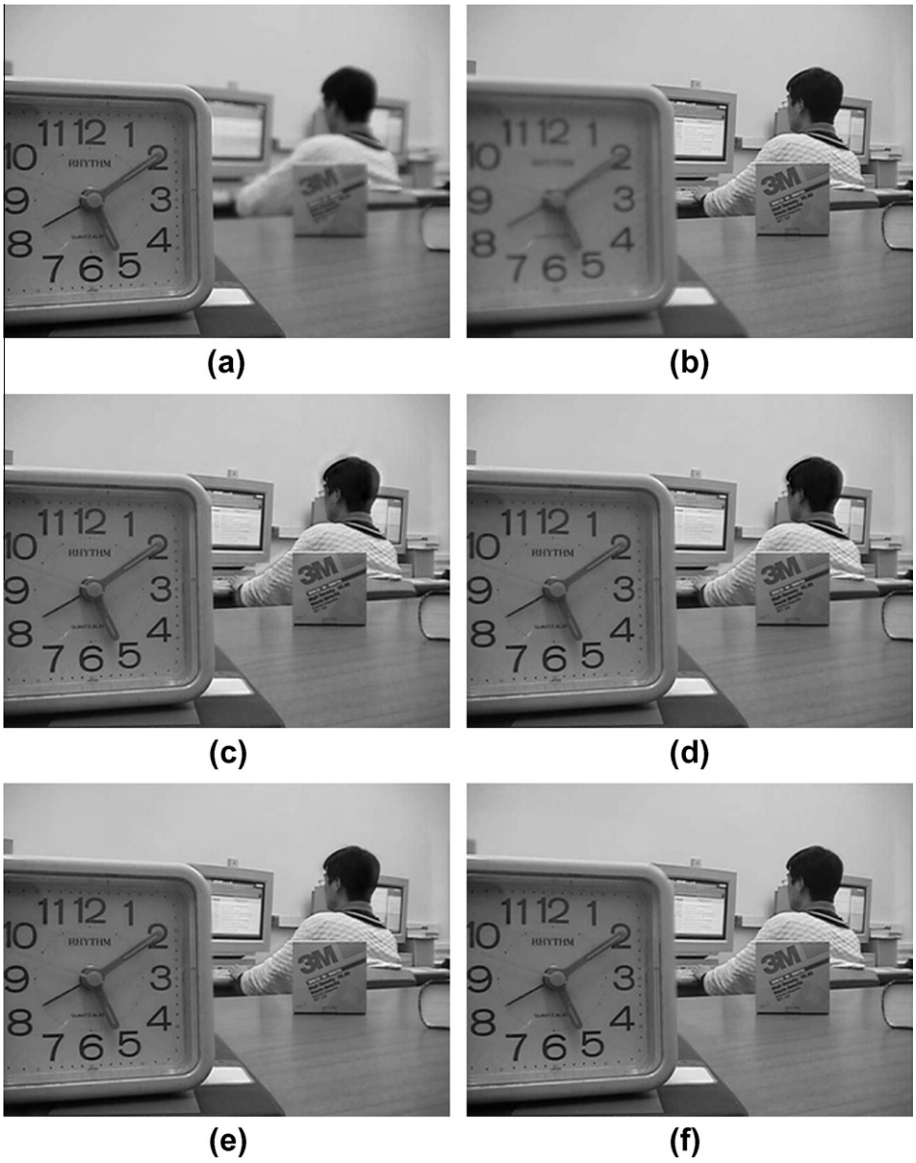

Fig. 13. The "lab" source images (a and b), the fused images by DTCWT (c), NSCT (d), SR (e), IG (f), RGSF (g), MM (h), proposed method (i), and reference fused image (j). (Image courtesy of the Signal Processing and Communication Research Lab of Lehigh University.).

The second experiment is performed on the "trunk" source images shown in Fig. 9a and b (trunk and background in focus, respectively). In this experiment, as the movement of the girl is very obvious, the source images cannot be exactly registered. The fused images of different methods and the reference fused image are displayed in Fig. 9c–j. For clearer comparison, the magnified regions of the fused images are presented in Fig. 10a–h. Despite the fact that most details have been merged by the DTCWT, NSCT, SR and MM based methods, more or less artifacts appear in the motion area such as the head of the girl (see Fig. 10a–c and f). The IG and RGSF based methods, in spite of introducing fewer artifacts in the region of the girl, cannot preserve all high frequency details near the boundary of the trunk as shown in Fig. 10d and e. In comparison, as presented in Fig. 10g, there are no obvious artifacts in the fused image obtained by our method. Moreover, the fused image obtained by our method is the closest to the reference fused image (see Fig. 10h).

The pictures shown in Fig. 11a and b (leaves and grasses in focus, respectively) are taken outside with a breeze puffing across the grasses. In this case, although the camera is fixed, the source images cannot be registered perfectly because the grasses and leaves have subtle motion. So it is used to show the robustness of the proposed method for images having subtle motion. The fusion results obtained by different methods and the reference fused image are presented in Fig. 11c–j. The fusion decision maps of different spatial domain methods and the reference decision map are displayed in Fig. 11k–o. From Fig. 11c and d, it can be seen that the textures of the grasses from different source images are confused with each other by DTCWT or NSCT based fusion methods. From the five decision maps depicted in Fig. 11k–o, it can be observed that the result of the proposed method is the most closest to the manual segmentation result. Note that the fused image of SR based method in Fig. 11e shows that the result obtained by sparse representation is also satisfactory for this type of mis-registered images where the motion range is very limited.

The multi-focus images in Fig. 12a and b (foliages and hydrant in focus, respectively) are registered perfectly. The fused images obtained by different methods and the reference fused image are

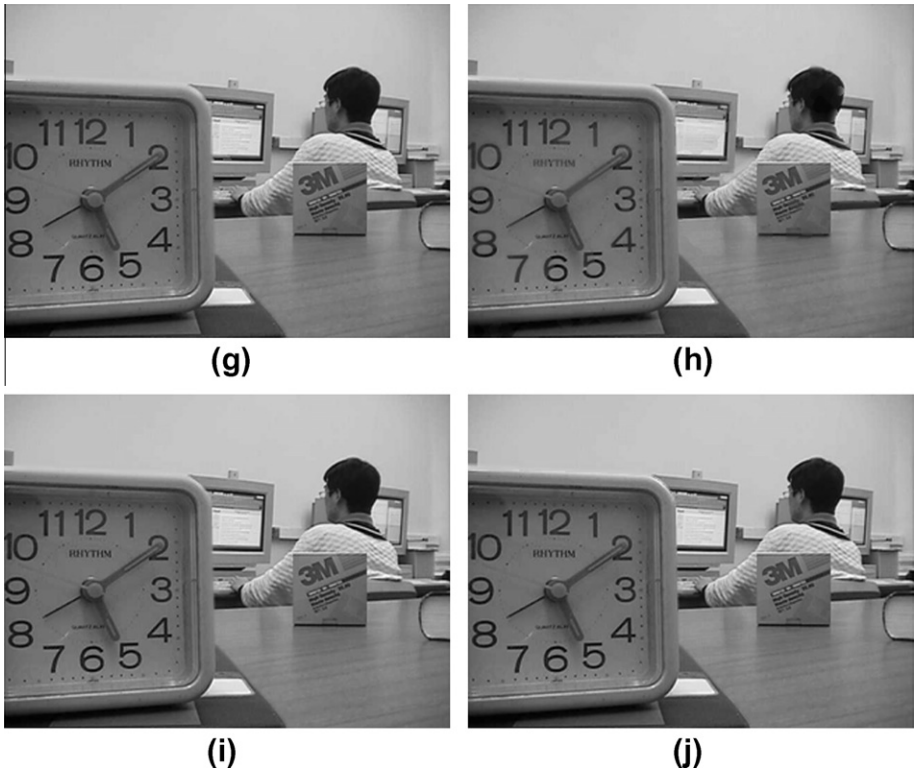

Fig. 13 (*continued*)

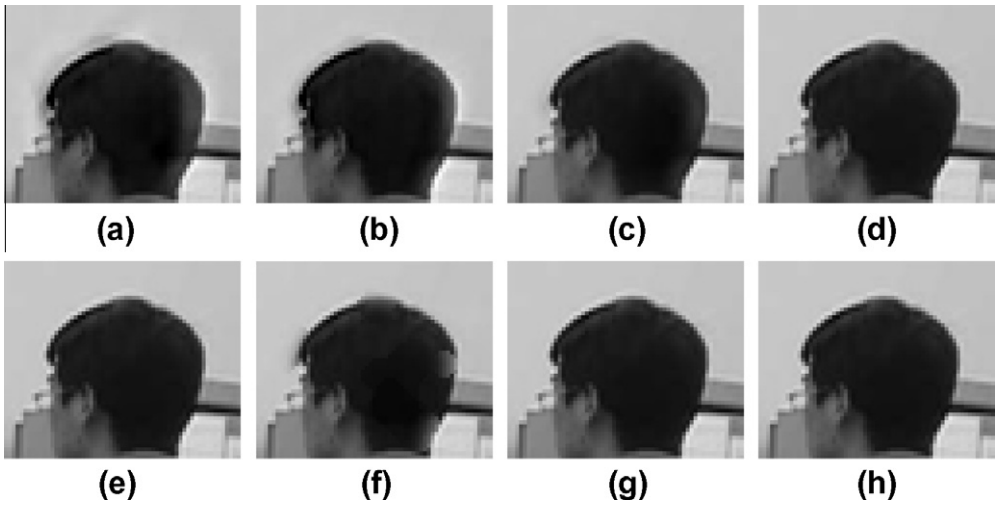

Fig. 14. Magnified regions of the fused images by DTCWT (a), NSCT (b), SR (c), IG (d), RGSF (e), MM (f), proposed method (g), and reference fused image (h).

presented in Fig. 12c–j. In addition, the fusion decision maps of spatial domain methods and the reference decision map are displayed in Fig. 12k–o. Because of the complex patterns (the flowers and the green foliages) of source images, IG, RGSF and MM based methods cannot obtain very accurate focused region, as illustrated in Fig. 12k–m. In comparison, since image matting is able to find very accurate outline of the focused object shown in Fig. 12n, the proposed method can obtain more satisfactory fused image which is presented in Fig. 12i.

In order to demonstrate that the proposed method can also work for standard test images, the experiment is also performed over one pair of gray multi-focus images, Fig. 13a and b (the clock and the background in focus, respectively). The fused images of different methods and the reference fused image are presented in Fig. 13c–j. For clearer comparison, magnified regions of the fused images are presented in Fig. 14a–h. It can be observed that the fused images obtained by DTCWT, NSCT, SR and MM based methods (see Fig. 14a–c and f) show artifacts on the head part of the man. Although the results of IG, RGSF (see Fig. 14d and e) are also encouraging in this area, they introduce artifacts in other parts. For instance, as shown in Fig. 13f and g, the IG based method cannot preserve the focused square part which is above the clock and the RGSF based method shows artifacts in the region between the clock and the computer screen.

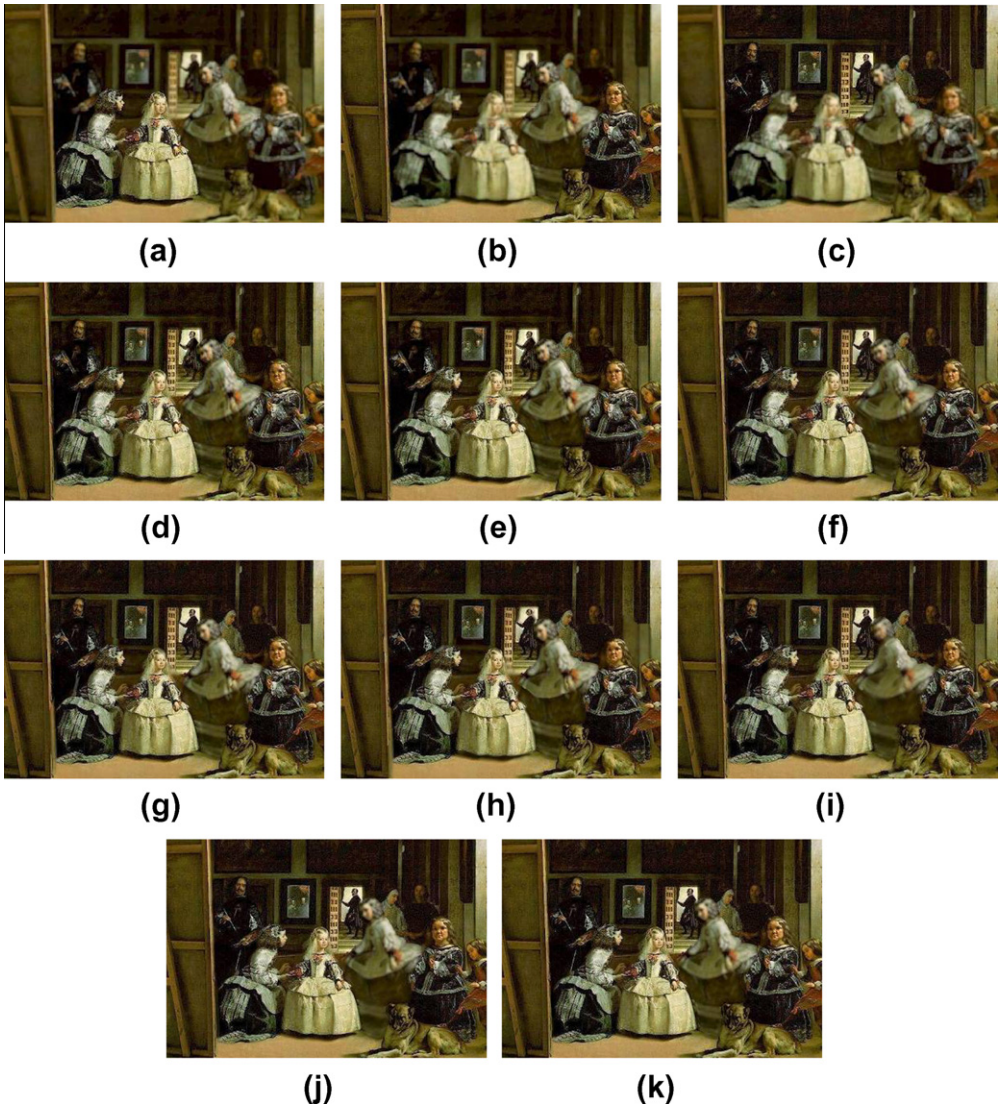

Fig. 15. The "oil painting" source images (a, b and c) and fused images DTCWT (d), NSCT (e), SR (f), IG (g), RGSF (h), MM (i), proposed method (j), reference (k).

Furthermore, we also perform the experiment over the "oil painting images" which have more than two source images, shown in Fig. 15a–c (two girls in the front, the front man on the right and the man on the left in focus, respectively). For three source mages, the proposed method merges the two of them at first. Then, the final fused image is obtained by merging the fused image obtained above with the left source image. The fused images of different methods are presented in Fig. 15d–j. Fig. 15k is the reference fused image. Through Fig. 15j, it can be seen that all important information of the source images has been integrated into our fused image and no artifact been introduced. Thus, the proposed method is also suitable for the case with more than two source images. Fig. 8–15 shows that the results from the proposed method are superior to all the other results visually when fusing multi-focus images in dynamic scenes, perfectly registered images with complex image patterns, standard test images and more than two source images. In the following, in order to assess the fusion performance more objectively, the modified mutual information (M_F^{XY}), similarity based quality index ($Q(X, Y, F)$), structural similarity index (SSIM), and gradient based index ($Q^{XY/F}$) are used to evaluate the fusion performance of different fusion methods. Table 1 shows the quantitative assessments of the fused images obtained by different fusion algorithms. From this table, it can be seen that our proposed method outperforms the DTCWT, NSCT, SR, IG, RGSF and MM based methods in terms of the SSIM, $Q^{XY/F}$ and $Q(X, Y, F)$, except that $Q(X, Y, F)$ in Figs. 8 and 15 are not the largest. It indicates that the fused images obtained by our method are the closest to the reference images. In addition, the degradation of edge and structure information introduced to the proposed fusion method is less than other algorithms. For the modified mutual information metric M_F^{XY}, it measures the statistical relationship between source images and the fused image. However, unlike the gradient-based metric $Q^{XY/F}$, it lacks in estimating the preservation of local structures of source images. This limitation leads to the questionable results of M_F^{XY} favoring IG and RGSF based methods over the proposed fusion algorithm. Even so, the values of M_F^{XY} of the fused images obtained by the proposed method are still acceptable, e.g., rank as the first in Fig. 15, the second in Figs. 11 and 13 and the third in Figs. 8, 9 and 12. Based on the analysis above, it can be concluded that both by visual comparison and by objective evaluation, the proposed method shows competitive fusion performance compared with previous methods.

Furthermore, in order to observe the dependence of the proposed algorithm on image noise, one experiment is performed over

Table 1
The quantitative assessments of different multi-focus image fusion methods.

Images	Indexes	Methods						
		DTCWT	NSCT	SR	IG	RGSF	MM	Our
Fig. 8	M_F^{XY}	0.3654	0.3705	0.3976	0.5928	**0.6204**	0.5861	0.5877
	SSIM	0.9154	0.9230	0.9478	0.8976	0.9061	0.9430	**0.9702**
	$Q(X, Y, F)$	0.9281	0.9335	0.9698	**0.9826**	0.9814	0.9673	0.9812
	$Q^{XY/F}$	0.6417	0.6489	0.7021	0.6937	0.6913	0.7052	**0.7086**
Fig. 9	M_F^{XY}	0.7556	0.7627	0.7827	1.0401	**1.0417**	0.9935	1.0331
	SSIM	0.9623	0.9651	0.9871	0.9921	0.9908	0.9825	**0.9965**
	$Q(X, Y, F)$	0.9480	0.9499	0.9642	0.9889	0.9676	0.9756	**0.9891**
	$Q^{XY/F}$	0.6198	0.6249	0.6763	0.6822	0.6784	0.6731	**0.6839**
Fig. 11	M_F^{XY}	0.5842	0.5795	0.6617	**0.8994**	0.8633	0.8736	0.8897
	SSIM	0.9341	0.9324	0.9752	0.9520	0.9649	0.9720	**0.9867**
	$Q(X, Y, F)$	0.9348	0.9319	0.9839	0.9872	0.9853	0.9811	**0.9882**
	$Q^{XY/F}$	0.5809	0.5823	0.6772	0.6776	0.6691	0.6738	**0.6802**
Fig. 12	M_F^{XY}	0.8617	0.8712	0.8757	**1.0461**	1.0437	0.9798	1.0305
	SSIM	0.9785	0.9801	0.9901	0.9890	0.9858	0.9777	**0.9953**
	$Q(X, Y, F)$	0.9336	0.9399	0.9636	0.9752	0.9768	0.9535	**0.9772**
	$Q^{XY/F}$	0.7108	0.7201	0.7351	0.7327	0.7299	0.7223	**0.7362**
Fig. 13	M_F^{XY}	0.8153	0.8235	0.8873	0.9061	**0.9245**	0.8694	0.9192
	SSIM	0.9850	0.9878	0.9932	0.9929	0.9969	0.9815	**0.9980**
	$Q(X, Y, F)$	0.9032	0.9131	0.9002	0.9622	0.9817	0.9186	**0.9821**
	$Q^{XY/F}$	0.7058	0.7149	0.7322	0.7422	0.7396	0.7119	**0.7424**
Fig. 15	M_F^{XY}	0.6733	0.6735	0.8654	0.9204	0.9328	0.9163	**0.9541**
	SSIM	0.8786	0.8790	0.9502	0.9445	0.9272	0.9590	**0.9817**
	$Q(X, Y, F)$	0.8900	0.8877	0.9721	**0.9892**	0.9884	0.9670	0.9855
	$Q^{XY/F}$	0.3895	0.3884	0.4410	0.4553	0.4418	0.4471	**0.4628**

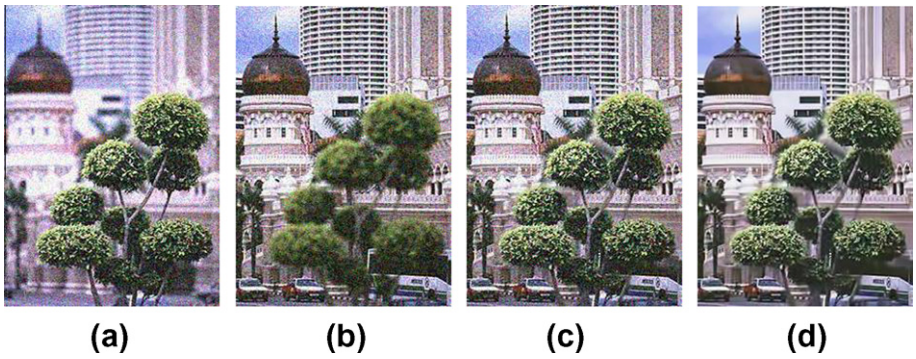

Fig. 16. The noisy "tree" source images (a and b), the fused image obtained by the proposed method (c), the final fusion result after denoising (d).

noisy images. Fig. 16a and b is the noisy multi-focus images corrupted by zero-mean Gaussian white noise with $\sigma = 30$. Using noisy images as inputs, the fusion result obtained by the proposed method is presented in Fig. 16c. It can be seen that the fused image is also satisfactory. For instance, the focused car and tree are both well preserved. However, since the fusion process of the proposed method is performed on spatial domain, the Gaussian noise cannot be removed in the fusion process. In order to remove these noises, a fast denoising method [36] is performed on Fig. 16c. Fig. 16d shows the de-noised fused image which has more satisfactory visual quality.

Finally, in order to show the potential of the proposed method in other applications of image fusion, we perform an experiment over a pair of infrared and visible images shown in Fig. 17a and b. It can be seen that warm objects stand out well against cool backgrounds in the infrared image, i.e., the person is easily visible against the environment, while the visible image reveals more visible details of the background such as the fence. For the infrared image, because the objects with higher luminance are usually more salient, the trimap generation process of our method should be modified. In this experiment, the definite "focused" region of the infrared image is obtained by binarization. The threshold used for binarization is set as three-quarters of the maximum gray value. Fig. 17c, d and e shows the fused images obtained by the DTCWT, NSCT and the proposed method, respectively. It can be seen that, in this case, our method can generate more satisfactory fused image (see Fig. 17e), i.e., the man and fence in Fig. 17e are clearer compared with those in Fig. 17c and d.

5. Conclusions

In this paper, a novel approach is proposed to fuse multi-focus images in dynamic scenes. Unlike previous fusion methods which compare the focus values of pixels or regions to decide where is in focus, our approach forwards the focus information to image matting to find the focused region. Since image matting is able to make full use of the strong correlations between nearby pixels,

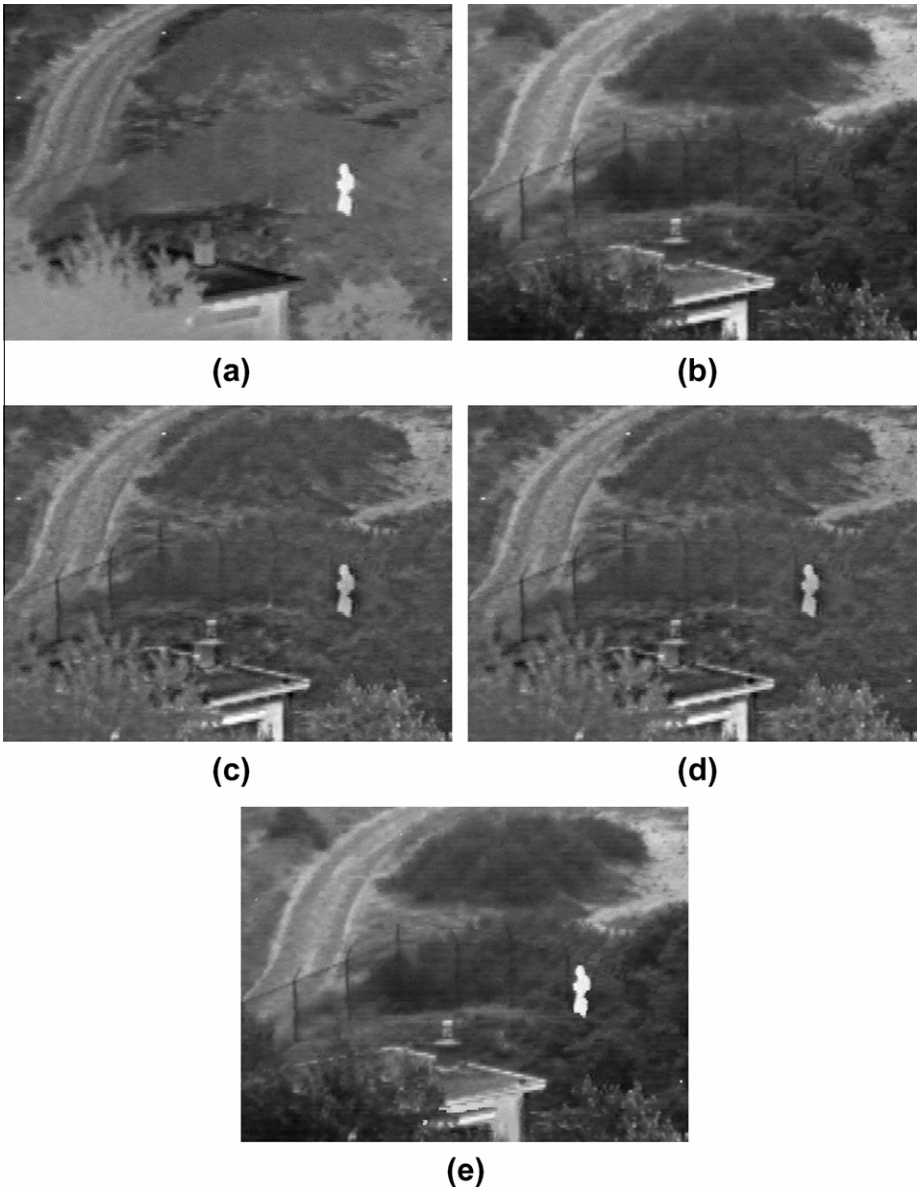

Fig. 17. The infrared and visible images (a and b), the fused images by DTCWT (c), NSCT (d), proposed method (e). (Image courtesy of Alexander Toet).

the proposed method shows better fusion performance compared with previous methods when merging multi-focus images in dynamic scenes. Besides, the robustness of the proposed method to image noise is demonstrated. The potential of the proposed method in infrared and visible image fusion is also presented. However, it is only a premature attempt. In the future, it is worthy to further investigate the effectiveness of the proposed method for other related tasks of image fusion.

Acknowledgments

The authors would like to thank the editor and anonymous reviewers for their detailed review, valuable comments and constructive suggestions. This paper is supported by the National Natural Science Foundation of China (Nos. 60871096 and 60835004), the Ph.D. Programs Foundation of Ministry of Education of China (No. 200805320006), the Key Project of Chinese Ministry of Education (2009-120), the Fundamental Research Funds for the Central Universities, Hunan University, and the Open Projects Program of National Laboratory of Pattern Recognition.

References

[1] A.A. Goshtasby, S. Nikolov, Image fusion: advances in the state of the art, Information Fusion 8 (2) (2007) 114–118.
[2] G. Pajares, J.M. de la Cruz, A wavelet-based image fusion tutorial, Pattern Recognition 37 (9) (2004) 1855–1872.
[3] V.S. Petrović, C.S. Xydeas, Gradient-based multiresolution image fusion, IEEE Transactions on Image Processing 13 (2) (2004) 228–237.
[4] J.J. Lewis, R.J. O'Callaghan, S.G. Nikolov, D.R. Bull, N. Canagarajah, Pixel- and region-based image fusion with complex wavelets, Information Fusion 8 (2) (2007) 119–130.
[5] E. Candès, L. Demanet, D. Donoho, L. Ying, Fast discrete curvelet transforms, SIAM Multiscale Modeling and Simulation 5 (3) (2006) 861–899.
[6] M.N. Do, M. Vetterli, The contourlet transform: an efficient directional multiresolution image representation, IEEE Transactions on Image Processing 14 (12) (2005) 2091–2106.
[7] Q. Zhang, B. Guo, Multi-focus image fusion using the nonsubsampled contourlet transform, Signal Processing 89 (7) (2009) 1334–1346.
[8] S. Yang, M. Wang, L. Jiao, R. Wu, Z. Wang, Image fusion based on a new contourlet packet, Information Fusion 11 (2) (2010) 78–84.

[9] T. Li, Y. Wang, Biological image fusion using a NSCT based variable-weight method, Information Fusion 12 (2) (2010) 85–92.
[10] R. Redondo, F. Šroubek, S. Fischer, G. Cristóbal, Multifocus image fusion using the log-Gabor transform and a multisize windows technique, Information Fusion 10 (2) (2009) 163–171.
[11] S. Yang, M. Wang, L. Jiao, Fusion of multispectral and panchromatic images based on support value transform and adaptive principal component analysis, Information Fusion 13 (3) (2012) 177–184.
[12] S. Daneshvar, H. Ghassemian, MRI and PET image fusion by combining IHS and retina-inspired models, Information Fusion 11 (2) (2010) 114–123.
[13] B. Yang, S. Li, Multifocus image fusion and restoration with sparse representation, IEEE Transactions on Instrumentation and Measurement 59 (4) (2010) 884–892.
[14] S. Li, J.T. Kwok, Y. Wang, Multifocus image fusion using artificial neural networks, Pattern Recognition Letters 23 (8) (2002) 985–997.
[15] Z. Wang, Y. Ma, J. Gu, Multi-focus image fusion using PCNN, Pattern Recognition 43 (6) (2010) 2003–2016.
[16] H.A. Eltoukhy, S. Kavusi, A computationally efficient algorithm for multi-focus image reconstruction, in: Proceedings of SPIE Electronic Imaging, San Jose, USA, 2003, pp. 332–341.
[17] S. Li, J.T. Kwok, Y. Wang, Combination of images with diverse focuses using the spatial frequency, Information Fusion 2 (3) (2001) 169–176.
[18] S. Li, B. Yang, Multifocus image fusion using region segmentation and spatial frequency, Image and Vision Computing 26 (7) (2008) 971–979.
[19] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, David Salesin, M. Cohen, Interactive digital photomontage, ACM Transactions on Graphics 23 (3) (2004) 292–300.
[20] M. Zribi, Non-parametric and region-based image fusion with Bootstrap sampling, Information Fusion 11 (2) (2010) 85–94.
[21] J. Wang, M.F. Cohen, Image and video matting: a survey, Foundations and Trends in Computer Graphics and Vision 3 (2) (2007) 97–175.
[22] H. Li, K.N. Ngan, Unsupervised video segmentation with low depth of field, IEEE Transactions on Circuits and Systems for Video Technology 17 (12) (2007) 1742–1751.
[23] Z. Liu, W. Li, L. Shen, Z. Han, Z. Zhang, Automatic segmentation of focused objects from images with low depth of field, Pattern Recognition Letters 31 (7) (2010) 572–581.
[24] J. Wang, M.F. Cohen, Optimized color sampling for robust matting, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, USA, 2007, pp. 1–8.
[25] L. Grady, Random walks for image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (11) (2006) 1768–1783.
[26] I. De, B. Chanda, B. Chattopadhyay, Enhancing effective depth-of field by image fusion using mathematical morphology, Image and Vision Computing 24 (12) (2006) 1278–1287.
[27] R.C. Gonzalez, R.E. Woods, S. Eddins, Digital image processing using MATLAB, Prentice Hall, New York, USA, 2004.
[28] G. Qu, D. Zhang, P. Yan, Information measure for performance of image fusion, Electronics Letters 38 (7) (2001) 313–315.
[29] C. Wei, R.S. Blum, Theoretical analysis of correlation-based quality measures for weighted averaging image fusion, Information Fusion 11 (4) (2010) 301–310.
[30] A. Toet, M.A. Hogervorst, S.G. Nikolov, J.J. Lewis, T.D. Dixon, D.R. Bull, C.N. Canagarajah, Towards cognitive image fusion, Information Fusion 11 (2) (2010) 95–113.
[31] T.D. Dixon, E.F. Canga, S.G. Nikolov, T. Troscianko, J.M. Noyes, C.N. Canagarajah, D.R. Bull, Selection of image fusion quality measures: objective, subjective, and metric assessment, Journal of The Optical Society of America A – Optics Image Science and Vision 24 (12) (2007) B125–B135.
[32] M. Hossny, S. Nahavandi, D. Creighton, Comments on 'Information measure for performance of image fusion', Electronics Letters 44 (18) (2008) 1066–1067.
[33] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612.
[34] C. Yang, J. Zhang, X. Wang, X. Liu, A novel similarity based quality metric for image fusion, Information Fusion 9 (2) (2008) 156–160.
[35] C.S. Xydeas, V.S. Petrović, Objective image fusion performance measure, Electronics Letters 36 (4) (2000) 308–309.
[36] K. Dabov, A. Foi, V. Katkovnik, K. Egiazarian, Color image denoising via sparse 3D collaborative filtering with grouping constraint in luminance–chrominance space, in: Proceedings of IEEE International Conference on Image Processing, San Antonio, USA, 2007, pp. I-313–I-316.