Information Fusion 25 (2015) 72–84

Contents lists available at ScienceDirect

Information Fusion

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

Multi-focus image fusion using dictionary-based sparse representation

Mansour Nejati [a], Shadrokh Samavi [a,b,*], Shahram Shirani [b]

[a] Department of Electrical and Computer Engineering, Isfahan University of Technology, Iran
[b] Department of Electrical and Computer Engineering, McMaster University, Hamilton, Canada

ARTICLE INFO

Article history:
Received 31 January 2014
Received in revised form 7 September 2014
Accepted 19 October 2014
Available online 1 November 2014

Keywords:
Multi-focus image fusion
Dictionary learning
K-SVD
Sparse representation
Guided image filtering

ABSTRACT

Multi-focus image fusion has emerged as a major topic in image processing to generate all-focus images with increased depth-of-field from multi-focus photographs. Different approaches have been used in spatial or transform domain for this purpose. But most of them are subject to one or more of image fusion quality degradations such as blocking artifacts, ringing effects, artificial edges, halo artifacts, contrast decrease, sharpness reduction, and misalignment of decision map with object boundaries. In this paper we present a novel multi-focus image fusion method in spatial domain that utilizes a dictionary which is learned from local patches of source images. Sparse representation of relative sharpness measure over this trained dictionary are pooled together to get the corresponding pooled features. Correlation of the pooled features with sparse representations of input images produces a pixel level score for decision map of fusion. Final regularized decision map is obtained using Markov Random Field (MRF) optimization. We also gathered a new color multi-focus image dataset which has more variety than traditional multi-focus image sets. Experimental results demonstrate that our proposed method outperforms existing state-of-the-art methods, in terms of visual and quantitative evaluations.

1. Introduction

Depth of field in optical lenses of conventional cameras is limited. Thus, only the objects at a particular distance from the camera are in focus and captured sharply whereas objects at other distances in front of or behind the focus plane are defocused and blurred. However, for accurately interpreting and analyzing images, it is desired to obtain images with every object in focus [1]. Multi-focus image fusion is an effective technique to solve this problem by combining two or more images of the same scene taken with different focus settings into a single all-in-focus image with extended depth of field, which is very useful for human or machine perception. The multi-focus image fusion has been applied in various applications such as microscopic imaging, remote sensing, and computer vision [1].

During the past years, many multi-focus image fusion algorithms have been developed [3–16]. According to the fusion domain, these algorithms could be categorized into two main groups: transform domain and spatial domain fusion [2]. In the first group, transform coefficients are fused and the fused image is reconstructed from these composite coefficients. The transform domain fusion methods that are based on multi-scale transforms are the most commonly used methods in this group [3]. Many kinds of multi-scale transforms have been proposed and adopted for image fusion such as pyramid decomposition [4], discrete wavelet transform (DWT) [5,6], dual-tree complex wavelet transform (DTCWT) [7], and discrete cosine harmonic wavelet transform (DCHWT) [8]. Recently developed multiscale geometry analysis tools with higher directional sensitivity than wavelets, such as shearlet transform [9], curvelet transform (CVT) [10], and nonsubsampled contourlet transform (NSCT) [11] are employed too. Also, some novel signal decomposition methods like robust principal component analysis (RPCA) [1] and sparse representation (SR) [12–14], are also applied to image fusion. The transform domain fusion methods have three common basic steps. First, the source images are decomposed to get the transform coefficients. The transform coefficients are then integrated according to a certain fusion rule. The fused image is finally constructed by applying the inverse transform on the fused coefficients [12].

Unlike the transform domain fusion methods, in spatial domain methods, fusion rules are directly applied to image pixels or image regions. In general, spatial domain methods can be classified into two groups of pixel based [3,15,16], and region based methods [17,18]. The main principle in these methods is selecting the pixels or regions with more clarity according to some image clarity measure, namely focus measure, to construct the fused image. Energy of Laplacian and spatial frequency are two typical focus

* Corresponding author at: Department of Electrical and Computer Engineering, Isfahan University of Technology, Iran.

http://dx.doi.org/10.1016/j.inffus.2014.10.004
1566-2535/

measures used to make a decision about the clarity of pixels or regions. The main drawbacks of these spatial domain fusion methods are misalignment of decision map with boundary of focused objects and wrong decision in sub-regions of the focused or defocused regions which produce undesirable artifacts in the final fused image. To mitigate these artifacts, some spatial techniques use the weighted average of pixel values for fusing the source images, instead of using binary decision [3,16]. Depending on weight construction method, these methods lead to halo artifacts near some edges, contrast decrease, and/or reduction of sharpness.

In this paper we introduce a new multi-focus image fusion method in spatial domain which produces spatially smooth and edge aligned decision map to accurately merge the in-focus regions of multi-focus source images into a single fused image. Our method relies on dictionary learning and sparse representation of focus features. Sparse representation has proven to be an extremely powerful tool for analyzing a large class of signals [19]. In addition to successful application of sparse representation in many classical signal processing problems, such as compression and denoising [20,21], sparse-representation-based techniques are increasingly attracting attention in computer vision area due to its state-of-the-art performance in many applications, such as image classification [22,23], face recognition [24] action recognition [19], and object tracking [25]. Basic observation in these applications is that despite the images (or their features) are naturally very high dimensional, the images in the same class usually lie on a low-dimensional subspace [26]. Therefore, given a dictionary of representative samples for the distribution, it is expected that there exists a sparse representation with respect to such a dictionary for a typical sample.

State-of-the-art performance of sparse representation in many computer vision tasks motivated us to use this powerful tool in a new multi-focus image fusion framework. In this framework, a dictionary is learned for sparse representation of focus features extracted from small patches of source images. Max pooling is then applied to the sparse representations from training samples to get the pooled features. Correlation of training pooled features with the sparse representations of input source images produces a pixel level score map which is employed for an initial decision about the clarity of each pixel. Final regularized decision map is obtained using Markov Random Field (MRF) optimization. Our approach is different from previous sparse-representation-based fusion methods [12–14] in two aspects. First, fused images in those mentioned references are constructed by using the inverse transform of the fused sparse coefficients while our method has the advantage of working in the spatial domain. Second, we use the sparse representation, which is obtained from a learned dictionary, for feature extraction and classification of pixels as focused or unfocused ones. To validate the effectiveness of our method, extensive experiments are conducted using two datasets under three objective quality metrics. Experimental results demonstrate that our proposed method outperforms existing state-of-the-art methods, in terms of visual and quantitative evaluations.

The rest of this paper is organized as follows. In Section 2, the signal sparse representation theory is a briefly reviewed. Section 3 describes the details of the proposed multi-focus image fusion based on dictionary learning and sparse representation of focus feature. Experimental results, comparison to state-of-the-arts and objective evaluations are demonstrated in Section 4. Finally, Section 5 concludes the paper.

2. Sparse representation theory

Sparse signal representations have recently drawn much interest in vision, signal and image processing [26], [27]. Sparse signal representation has proven to be an extremely powerful tool for analyzing a large class of signals. This is mainly due to the fact that signals and images of interest can be sparse or compressible in some dictionary of bases [3]. In sparse representation modeling of an input signal $y \in \mathbb{R}^n$, it is represented as a linear combination of a few atoms of an over-complete dictionary $\Phi \in \mathbb{R}^{n\times K}$ $(K > n)$ as

$$y = \Phi x \tag{1}$$

where the vector $x \in \mathbb{R}^K$ contains the representation coefficients of the signal y and the dictionary matrix Φ contains K prototype signals referred as atoms for columns, $\{\phi_j\}_{j=1}^K$. The linear system in (1) with a full-rank over-complete dictionary Φ, becomes an underdetermined system of linear equations having an infinite number of solutions, hence constraints on the solution must be set. Finding the sparsest solution with the fewest number of non-zero coefficients involves solving the optimization problem

$$\min_x \|x\|_0 \quad \text{subject to} \quad y = \Phi x \tag{2}$$

in exact representation, or

$$\min_x \|x\|_0 \quad \text{subject to} \quad \|y - \Phi x\|_2 \leqslant \epsilon \tag{3}$$

in approximate representation [20], where $\|\cdot\|_0$ is the ℓ_0 semi-norm that counts the number of nonzero entries in a vector. These are in general NP-hard problems and thus, approximation techniques such as pursuit algorithms are used to get an approximated solution [20]. Orthogonal matching pursuit (OMP) is a greedy pursuit algorithm which is widely used because of its simplicity and efficiency [21].

Construction of a proper dictionary is an important issue in sparse representation. It has been observed that learning a dictionary directly from training signals leads to better representation and hence provides superior performance when compared to predefined dictionaries (such as Fourier or wavelet) in many image and vision applications [28]. Particularly, in computer vision, we often have to learn a task-specific dictionary from given sample images [26].

Methods for learning a dictionary for sparse coding from the training data have been proposed recently [20,23,29]. Let $Y \in \mathbb{R}^{n\times N}$ be a set of N input signals of dimension n, *i.e.* $Y = \{y_i\}_{i=1}^N, y_i \in \mathbb{R}^n$. Learning a dictionary $\Phi \in \mathbb{R}^{n\times K}$ $(K > n)$ with K atoms for sparse representation of Y is formally written as the following optimization problem:

$$\hat{\Phi}, \hat{X} = \underset{\Phi, X}{\operatorname{argmin}} \|Y - \Phi X\|_F^2 \quad \text{subject to} \quad \forall i, \|x_i\|_0 \leqslant T \tag{4}$$

where $\|\cdot\|_F$ is the Frobenius norm, $X = \{x_i\}_{i=1}^N, x_i \in \mathbb{R}^K$ are the sparse representation of input signals Y, and T is a sparsity constraint of sparse representations to be contained no more than T nonzero coefficients.

3. Proposed multi-focus image fusion

Motivated by the powerful ability of sparse coding in classification, we propose a new framework for multi-focus image fusion base on learning dictionary and its corresponding sparse representation using focus measure of local patches of source images. In this framework, every pixel from each source image is classified as an either in-focus pixel to be used in the fused image or not. According to the classification, a decision map for fusing of input images is obtained that label of each pixel in this map indicates intensity value of which input images must be used for that location in fused image.

The proposed algorithm consists of two phases: training and testing. Fig. 1 gives an overview of the proposed method for the case of two source images. In the training phase, an overcomplete

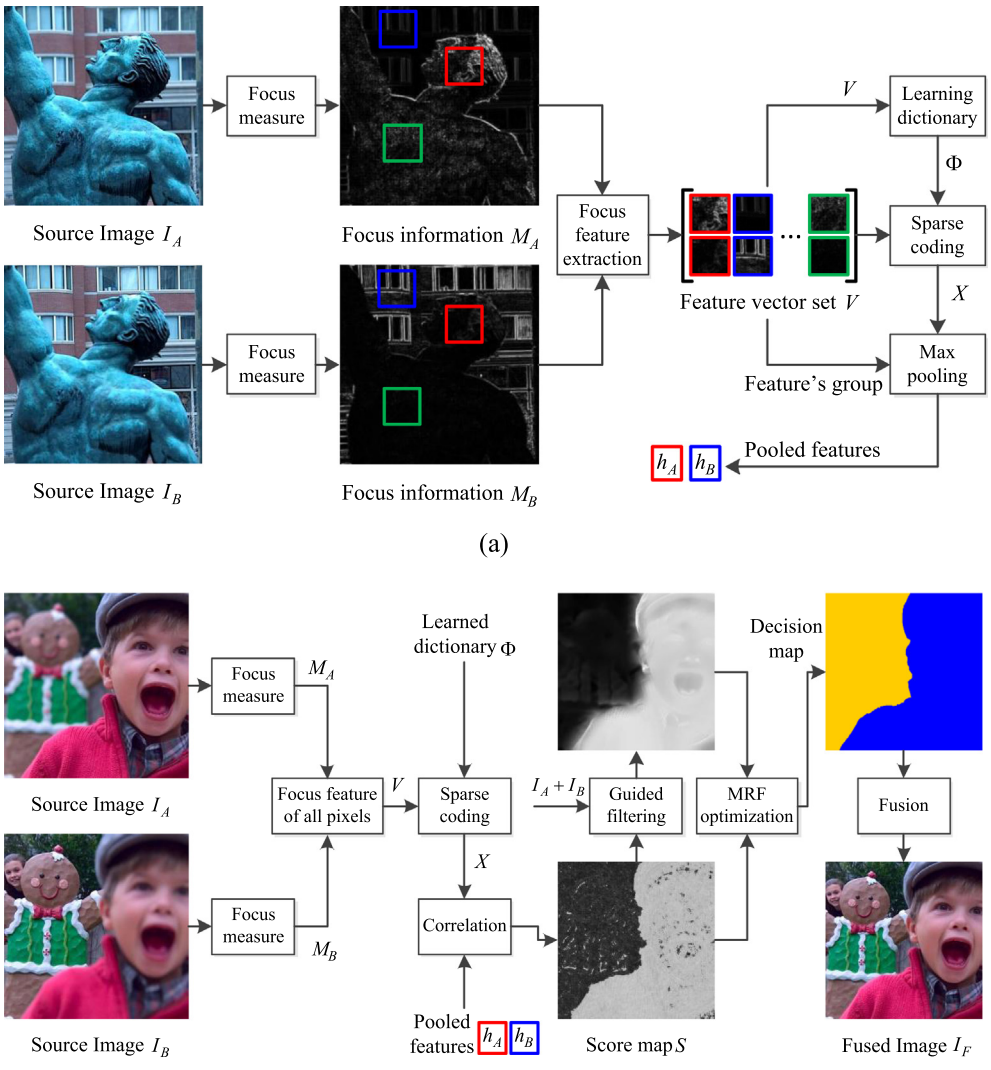


Fig. 1. Overview of the proposed multi-focus image fusion method. (a) Training phase. (b) Testing phase.

dictionary is learned from focus information map of source image patches using K-SVD algorithm [20]. We compute sparse representation of the training patches using the learned dictionary, and aggregate them by max pooling. Aggregation of sparse representations is accomplished in a supervised manner in which two pooled features corresponding to sparse representations that suggest the first source image is in focus or the second one, are obtained.

In the testing phase, focus information map of source images are computed. For each pixel location, the sparse representation of the two corresponding local patches in these two maps is obtained and its correlation to training pooled features is calculated. Then, the label of training pooled feature with larger correlation value is assigned to that pixel location in the decision map. Finally, to obtain a smooth decision map, where discontinuities are aligned with the source image edges, we use Markov Random Field (MRF) optimization and seek a regularized decision map labeling. Our approach can easily be applied to more than two source images which will be demonstrated in Section 4.3. Following subsections describe the above mentioned steps in detail.

3.1. Learning a dictionary

Fusion rule in spatial domain fusion methods is generally based on the comparison of some spatially computed focus or sharpness measure between source images in pixel/region level. Then the fusion rule selects the pixel/region with larger focus measure to construct the fused image.

In our algorithm we want to learn a dictionary such that we are able to make such a decision based on sparse representations. Thus for learning the relative degree of focus between source images, we use the focus information map of corresponding patches in two source images together as training samples. Focus information map of each source image is obtained by performing of a focus measure on it by using a sliding window technique. Several focus measures have been devised and applied to multi-focus image fusion such as energy of image gradient (EOG), variance of image intensities (VOI), energy of Laplacian of the image (EOL), and spatial frequency (SF) [30]. According to objective assessments in [30], energy of Laplacian is the preferred focus measure in spatial domain. For each pixel, EOL is computed as local average of the squared image Laplacian. Therefore, for source images I_A and I_B, the focus information maps are obtained as follows

$$M_q = H \otimes \left(\nabla^2 I_q\right)^2, \quad q \in \{A, B\} \tag{5}$$

where H is the averaging filter, $\otimes$ and ∇^2 denote convolution and Laplacian operators, respectively. For the filter H, we use a Gaussian filter of size 5×5 pixel in order to emphasize on Laplacian value of

pixels near to the center of local averaging window more than pixels far from the center. Laplacian is the simplest isotropic derivative operator which for a digital function (image) I of two variables i and j is given by

$$\nabla^2 I = \frac{\partial^2 I}{\partial i^2} + \frac{\partial^2 I}{\partial j^2} = [-I(i-1,\ j-1) - 4I(i-1,\ j) - I(i-1,\ j+1) - 4I(i,\ j-1) + 20I(i,\ j) - 4I(i,\ j+1) - I(i+1,\ j-1) - 4I(i+1,\ j) - I(i+1,\ j+1)] \quad (6)$$

Focus information map indicates the high frequency information of the source image and has small values in blurred areas. Two sample source images and their corresponding focus information maps are shown in Fig. 2.

After computing focus information maps we need to create training samples for dictionary learning. To do this, for each pair of training source images, a set of corresponding patch pairs are randomly sampled from their focus information maps (see Fig. 1(a)). All patches are then rearranged into vectors and the vectors of each two corresponding patches are concatenated to form our focus feature vectors $v_{ff}^i, i = 1, \ldots, N$ where N is the number of training features. The feature vectors are normalized by their L2 norm. With patches of size $p \times p$, a focus feature vector will be of dimension $n = 2 \times p \times p$.

To learn a dictionary from a set of focus features, $V = \{v_{ff}^i\}_{i=1}^N$, a recently developed K-SVD dictionary learning algorithm is used [20]. K-SVD is a standard unsupervised dictionary learning algorithm that iteratively solves the optimization problem (4) by alternating between computing X in sparse coding step, and Φ in dictionary update step. In the sparse coding step, Φ is kept fixed and X is efficiently computed using the greedy Orthogonal Matching Pursuit (OMP). Given the sparse representations X, the dictionary is updated sequentially by singular value decomposition. An example of a dictionary learned on a set of focus features extracted from 9×9 patches of focus information maps is visualized in Fig. 3.

3.2. Aggregation of sparse representations

Once the dictionary Φ is learned, we again use OMP to compute sparse representations $X = \{x^i\}_{i=1}^N, x^i \in \mathbb{R}^K$ of all training focus features $v_{ff}^i, i = 1, \ldots, N$. The focus features are grouped with supervision into two classes according to whether their corresponding patches are focused in source image I_A, or I_B. In fact, a selection mask is manually created for each pair of training source images in order to obtain a set of labeled training examples. This mask roughly specifies regions by labeling which location is focused in I_A and which one is focused in I_B. The patch pairs, used in the formation of training focus feature vectors, are randomly selected from the specified regions of the selection mask. Therefore, each training focus feature v_{ff}^i is assigned to one of the two classes using this selection mask. Accordingly, each sparse representation x^i belongs to one of these two classes. Let $X_A = \{x_A^i\}_{i=1}^{N_A}$ and $X_B = \{x_B^i\}_{i=1}^{N_B}$ with $N_A + N_B = N$ denote the sparse representations corresponding to class A and class B, respectively.

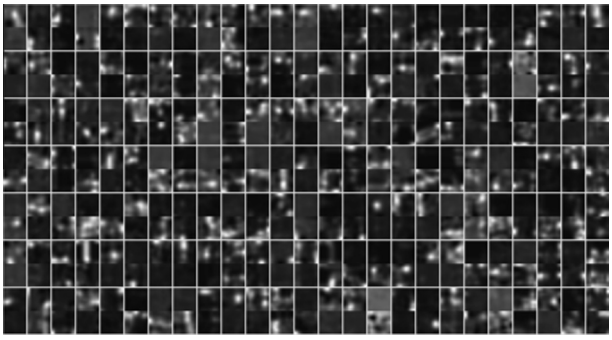


Fig. 3. An example of a learned dictionary using KSVD on focus features extracted from the "book" source images. 175 dictionary elements are shown which were randomly selected from 400 elements.

The sparse coefficients in each vector $x^i \in X$ give the contribution of all the dictionary atoms in approximating the focus feature $v_{ff}^i \in V$. The contribution of the dictionary atoms toward the representation of a particular class is thus collectively presented by sparse coefficients associated with all focus features of that class. Therefore, each class can be characterized by aggregating some statistics of corresponding sparse representations. To do this, pooling schemes is used. Pooling operators are used in many modern visual recognition algorithms to summarize the coded features over larger neighborhoods and achieve richer representations [31]. We apply max pooling on sparse coefficients in X_A and X_B to get the corresponding pooled coefficients h_A and h_B, respectively:

$$h_c(j) = \max\{|x_c^1(j)|, |x_c^2(j)|, \ldots, |x_c^{N_c}(j)|\}, \quad c \in \{A, B\} \quad (7)$$

where $h(j)$ is the j-th element of h, $x^i(j)$ is the j-th element in i-th sparse representation and N_c is the number of sparse representations that are pooled. One characteristics of sparse representation is that most dictionary atoms are rarely active across inputs. This attribute means that a pooling scheme that preserves more information for rare features will work well with sparse representation [32]. It has been shown that max pooling outperforms average pooling in many classification algorithms and it is also particularly well suited for the separation of features that are very sparse [33,34]. The pooled sparse representations are then used in test phase to produce a pixel level score value and decision map for fusion. Overview of the training step of our algorithm is presented in Fig. 1(a).

3.3. Pixel level classification using sparse representations

After learning a dictionary Φ and computing of pooled features h_A and h_B in training phase, we use them in multi-focus fusion of test images. For two multi-focus source images to be fused, we first compute focus information maps using (5). Starting from the

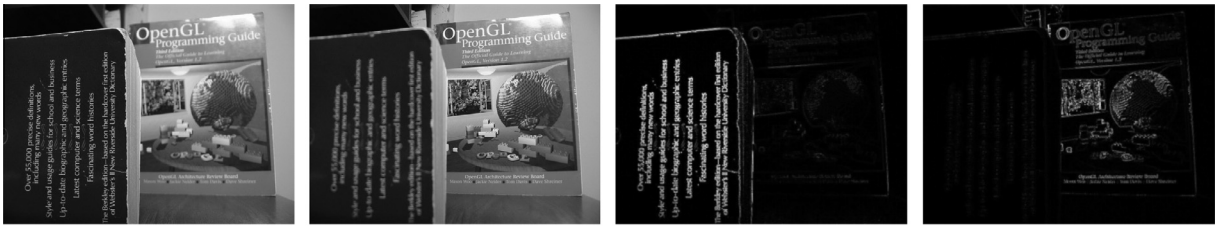
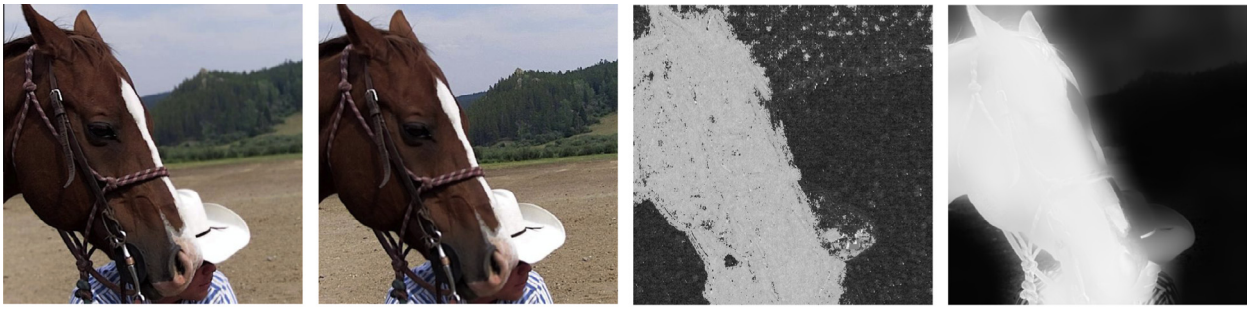


Fig. 2. The "book" source images (a and b), and their focus information maps (c and d).



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

Fig. 4. Guided filtering based refinement of score map. (a) and (b) are two input source images. (c) Score map obtained based on correlation of focus feature sparse representations and training pooled features. (d) Refined edge-aligned score map using guided filter with the average of source images serving as the guidance image.

top-left corner of the focus information maps, a sliding window, of size equal to patch size in the training phase, moves pixel by pixel and in each pixel location (i,j). Focus feature vector $v_{i,j}$ is formed in the same way as described in Section 3.1. Sparse representation of the focus feature vector over the learned dictionary Φ is then computed using the sparsity-constrained OMP such that each focus feature $v_{i,j}$ has fewer than T non-zero items in its decomposition $x_{i,j}$. Let $h_{i,j}$ denotes the absolute value of sparse coefficients. The correlation value of $h_{i,j}$ with pooled sparse coefficients h_A and h_B are used for classifying source image pixels as focused or defocused ones and create an initial decision map D of the same size of the input images:

$$D(i,j) = \begin{cases} 1 & if \quad h_{i,j}^T h_A > h_{i,j}^T h_B \\ 0 & if \quad h_{i,j}^T h_A \leqslant h_{i,j}^T h_B \end{cases} \tag{8}$$

where $D(i,j) = 1$ denotes that fused pixel in location (i, j) comes from the first source image I_A and $D(i,j) = 0$ denotes that fused pixel in location (i,j) comes from the second source image I_B. In addition to decision map D, a score map S is also obtained based on correlation values.

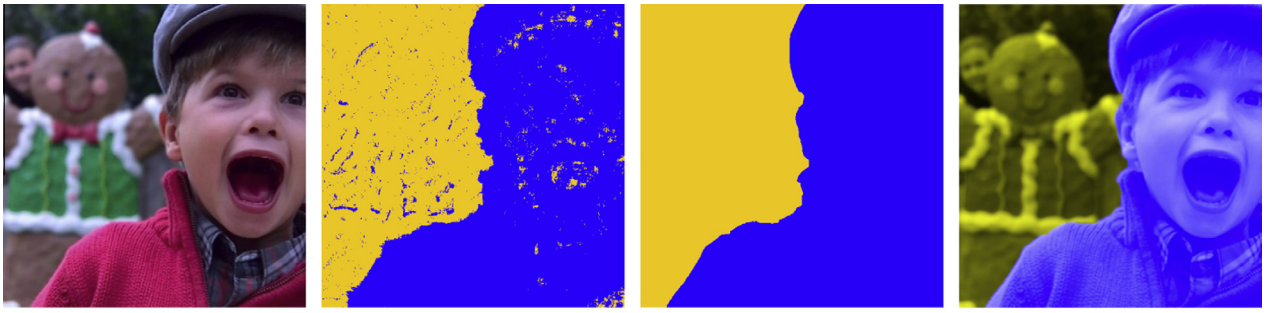
$$S(i,j) = \begin{cases} h_{i,j}^T h_A & if \quad D(i,\ j) = 1 \\ -h_{i,j}^T h_B & if \quad D(i,\ j) = 0 \end{cases} \tag{9}$$

Larger magnitudes of positive or negative scores indicate more confidence in the fusing decision. In fact, the absolute value of score map in each pixel position represents the degree of confidence and reliability of assigned label to that location in decision map D. But for distinguishing between confidence values of pixels with different labels (i.e. different decision values), we negated the case of $D(i,j) = 0$. This makes the obtained score map appropriate to be used in refinement of the decision map as it will be discussed in Section 3.4. Dividing the positive scores by the maximum positive one and dividing the negative scores by the minimum negative one produce normalize score values in the range of $[-1, 1]$. In the rest of paper, when mentioning score map, the normalized map is intended. Refinement of the decision map in final step of our algorithm is accomplished based on S.

3.4. Decision map regularization

Estimation of decision map D based on the correlation of sparse features with training ones may yield a non-smooth map with incorrect noisy labels in it. This is mainly due to flat (homogenous) regions that are similar in both focused and defocused states where high frequency information is not available and focus measures cannot differentiate them. Another reasons for incorrect labels in decision map, D, arise from presence of image noise and/or artifacts resulted from image compression. These add spurious and extraneous information, thus resulting in ambiguous selection in determining focused versus defocused pixels or regions [1].

We seek a regularized decision map labeling $\overline{D}$ which will be close to the initial estimate in (8), but will also be smooth. We also need the decision map discontinuities to be aligned with the image edges. But because of neighborhood processing in all the spatial focus measures, misalignments usually happen in transition edges between focused and defocused regions. Particularly, in spatial domain fusion methods those edge misalignments of the decision map with image edges produce visible artifacts around the transition boundaries of focused and defocused regions. The problem of decision map refinement can be viewed as a binary labeling problem where pixel labels indicate how the two source images should be fused into one. A popular approach to pixel labeling problems, which involve pixel interactions, is to construct a pairwise Markov Random Field (MRF) over image pixels and then estimate the MAP



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

Fig. 5. Regularizing decision map. (a) One of the source images with foreground in focus, (b) initial decision map, (c) decision map after our regularization approach, (d) overlay of (c) on (a).

solution of MRF model. Such pixel labeling problems are naturally formulated in terms of energy minimization, where the energy comprises a *data term* and a *smoothness term*:

$$E(\bar{D}) = \sum_{(i,j)} E_1(\bar{d}_{ij}) + \lambda \sum_{(i,j)} \sum_{(i',j') \in \mathcal{N}(i,j)} E_2(\bar{d}_{ij}, \bar{d}_{i'j'}) \tag{10}$$

where $\bar{d}_{ij}$ is the label of pixel location (i,j) in the refined decision map $\bar{D}(i,j)$, and $\mathcal{N}(i,j)$ is the set of neighbors of pixel (i,j), which is often the four nearest neighbors. For the data term, E_1, we use the quantized score map in order to put more energy cost on pixels with low scores:

$$E_1(\bar{d}_{ij}) = \begin{cases} S_Q(i,j) & \bar{d}_{ij} = d_{ij} \\ 1 & \bar{d}_{ij} \neq d_{ij} \end{cases} \tag{11}$$

Here, S_Q is obtained from score map S as follows

$$S_Q(i,j) = \begin{cases} 3 & |S(i,j)| < 0.2 \\ 2 & 0.2 \leqslant |S(i,j)| < 0.3 \\ 0 & |S(i,j)| > 0.3 \end{cases} \tag{12}$$

For smoothness energy term E_2, the score map S cannot be directly used. This is because of the misalignment of score map with object boundaries. To cope with this problem we use the guided filtering [35]. The guided filter is a recently proposed edge-preserving filter which computes the filtering output by considering the content of a guidance image and has a linear running time independent of the filter size. This filter is more generic than "smoothing" and it can transfer the structures of the guidance image to the filtering output which makes it qualified for applications such as image matting and dehazing [35]. In our algorithm, guided filter is used for smoothing the score map S, while it transfers the strong structure of source images to this map and generates a refined score map $\bar{S}$. For this purpose, the guided filter operates on the score map S as the input image and also takes the average of source images as the guidance image.

For instance, Fig. 4(d) shows the refined score map $\bar{S}$ obtained by guided filtering of initial score map shown in Fig. 4(c). The guidance image in this example is average of source images in Fig. 4(a) and Fig. 4(b). As shown in Fig. 4(c), guided filtering leads to a smooth and edge-aligned score map suitable for smoothness term of (10). The smoothness energy term between neighboring pixels is defined in order to persuade spatial consistency of the decision map:

$$E_2(\bar{d}_{ij}, \bar{d}_{i'j'}) = \begin{cases} 0 & \bar{d}_{ij} = \bar{d}_{i'j'} \\ \exp\left(-\frac{\|\bar{S}(i,j) - \bar{S}(i',j')\|_2}{\sigma^2}\right) & \bar{d}_{ij} \neq \bar{d}_{i'j'} \end{cases} \tag{13}$$

where $(i',j') \in \mathcal{N}(i,j)$ belongs to four nearest neighbors of pixel (i,j), and $\bar{S}$ is the refined score map based on guided filtering. The parameter σ controls the range of score value differences between two neighboring pixels in order to determine whether these two pixels should get the same label or should be labeled differently. The interaction energy function E_2 penalizes labeling discontinuities of neighboring pixels inversely proportional to the difference between their refined score map values. Therefore, the neighboring pixels with different labels which have similar score map values are penalized to get one label. In contrast, in boundaries of focused/unfocused regions where the difference $\|\bar{S}(i,j) - \bar{S}(i',j')\|_2$ is large for two neighboring pixels, the penalty becomes small.

We use the expansion-move algorithm based on graph cuts [36] to efficiently minimize the energy function defined in (10). This algorithm works by repeatedly calculating the global minimum of a binary labeling problem in its inner loop using graph cuts. At a given iteration, the expansion-move algorithm finds an optimal

Fig. 6. The grayscale image dataset including 8 pairs of grayscale multi-focus images.

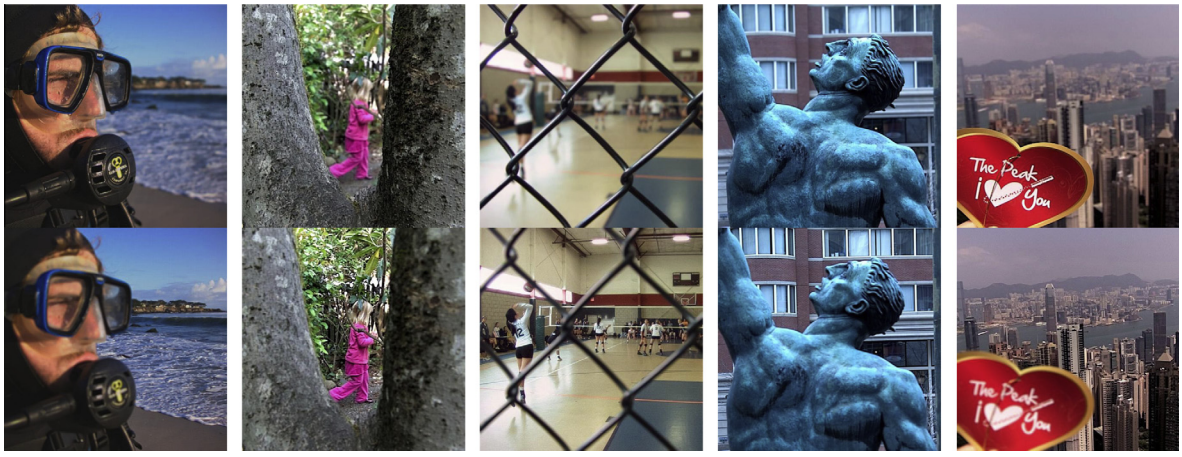


Fig. 7. Five sample image pairs from Lytro image dataset composed by twenty color multi-focus image pairs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 8. (a) The "book" source images and (b) the "toy" source images are served as training sources. (c) The learned dictionary on a collection of 9×9 patches sampled from focus information map of the training sources.

subset of pixels with fixed label α which would result in the largest decrease in the energy. The process repeatedly is done through all possible labels and results in a strong local minimum in the sense that no further improvement would be possible. Given the refined decision map $\overline{D}(i,j)$, the fused image I_F can be calculated by

$$I_F(i,j) = \overline{D}(i,j)I_A(i,j) + (1 - \overline{D}(i,j))I_B(i,j) \tag{14}$$

An example of decision map regularization is demonstrated in Fig. 5. It can be seen that the initial decision map shown in Fig. 5(b) is noisy and not aligned with object boundaries. As shown in Fig. 5(c), decision map regularization yields a smooth and edge-aligned labeling which is desirable for multi-focus image fusion.

4. Experiments

We evaluate the proposed multi-focus image fusion method on two image datasets. First one is the grayscale multi-focus dataset which contains 8 pairs of common grayscale images of different size used in many related papers. The second one is the new Lytro image dataset composed by 20 pairs of color multi-focus images of size 520×520 pixels which we have collected them from Lytro picture gallery[1] and is publicly available online [37]. Fig. 6 shows the grayscale multi-focus dataset. Furthermore, Fig. 7 shows some multi-focus image pairs of the Lytro image dataset.

The fusion results of the proposed method are compared with five recently proposed fusion algorithms. The first one is the multi-focus image fusion based on image matting (IFM) [3] which uses matting technique to obtain the accurate focused region of source images. The second compared algorithm is the guided filtering based fusion method (GFF) [38] in which the source images are decomposed into base and detail layers and weighted average of them is combined into fused image. The third method that is compared with our algorithm is the image fusion by wavelet-based statistical sharpness measure (WSSM) [5] that employs the spreading of the wavelet coefficients distribution as sharpness measure using a Laplacian mixture model. The fourth compared method is the cross bilateral filter-based image fusion method (CBF) [16] in which the fusing weights computed based on cross bilateral filtering of source images. Finally, the fifth compared method is the image fusion algorithm based on multi-scale discrete cosine harmonic wavelet transform (DCHWT) [8].

[1] https://pictures.lytro.com

4.1. Parameters setting

In our algorithm, 9×9 patches are used in computation of focus feature $v \in \mathbb{R}^n$ for the training samples and for each pixel location of the source images. Hence, each focus feature has dimension $n = 162$ which equals to dimension of each dictionary atom. Larger patches increase the computational cost of sparse coding and since decision about each pixel is taken based on a large spatial neighborhood, misalignments of decision map and objects boundaries become worse. Small patches on the other hand, lead to very noisy labeling in the decision map as it is more probable that small patches contain flat and not distinguishable regions. Patch sizes of 8 or 9 have been shown to be appropriate settings for the sparse representation-based image processing applications [20,21].

For the dictionary size, we set $K = 400$ and randomly chose 2000 patches to generate the dictionary for each dataset using K-SVD algorithm. We also fixed the sparsity factor $T = 5$ in all of our experiments. We tried other settings for the dictionary size and sparsity factor. With the larger ones, computational time increases while any substantial performance improvements is not observed. On the other hand, smaller dictionary sizes degrade the fusion performance.

The guided filter which is employed for refinement of score map, has two parameters: radius of filter, r, and regularization parameter, ϵ. The regularization parameter in the guided filter has similar effect as the range variance in the bilateral filter by which we can regulate the edge-preserving behavior of the filter. The smaller ϵ, the higher is the edge-preserving behavior. Since we want to transfer the structure of objects into score map and to have smooth regions in it, thus large filter size and small regularization parameter are preferred for score map refinement. We found that filter radius of $r = 30$–40 and regularization parameter $\epsilon = 10^{-3}$ are appropriate settings in our algorithm.

4.2. Objective image fusion quality metrics

For objective evaluation of fusion results, three fusion quality metrics including normalized mutual information (NMI) [39], Petrovic's metric $Q^{AB/F}$ [40], and visual information fidelity for fusion (VIFF) [41] are utilized in our work. The default parameters setting of these quality metrics given in the related publications are adopted.

(1) Normalized mutual information (NMI) [39] is an information theory based metric and a modified version of traditional mutual information, defined as:

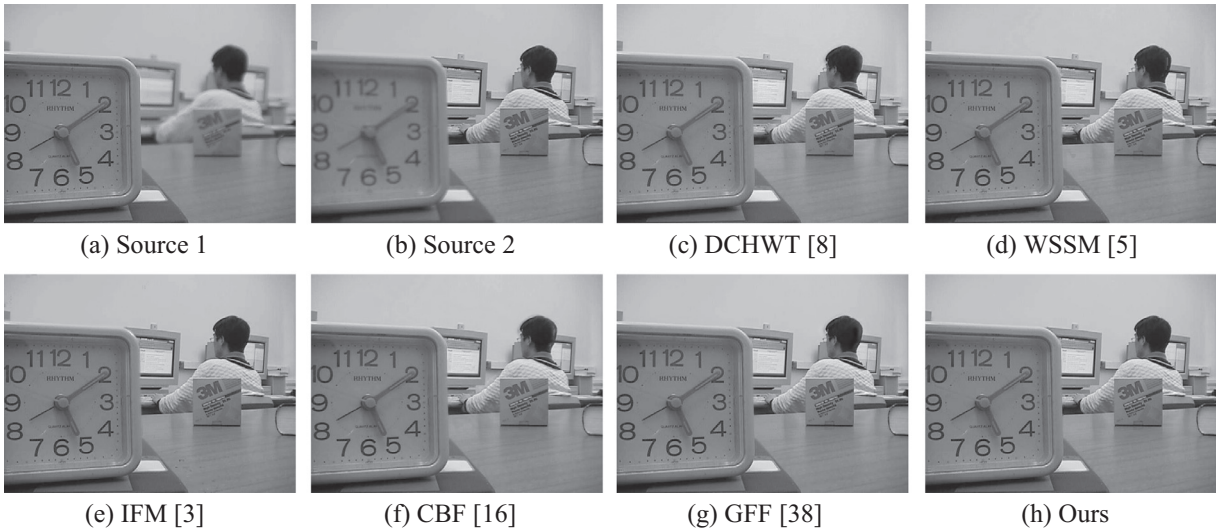


Fig. 9. The "lab" source images and fusion results obtained by different fusion methods.

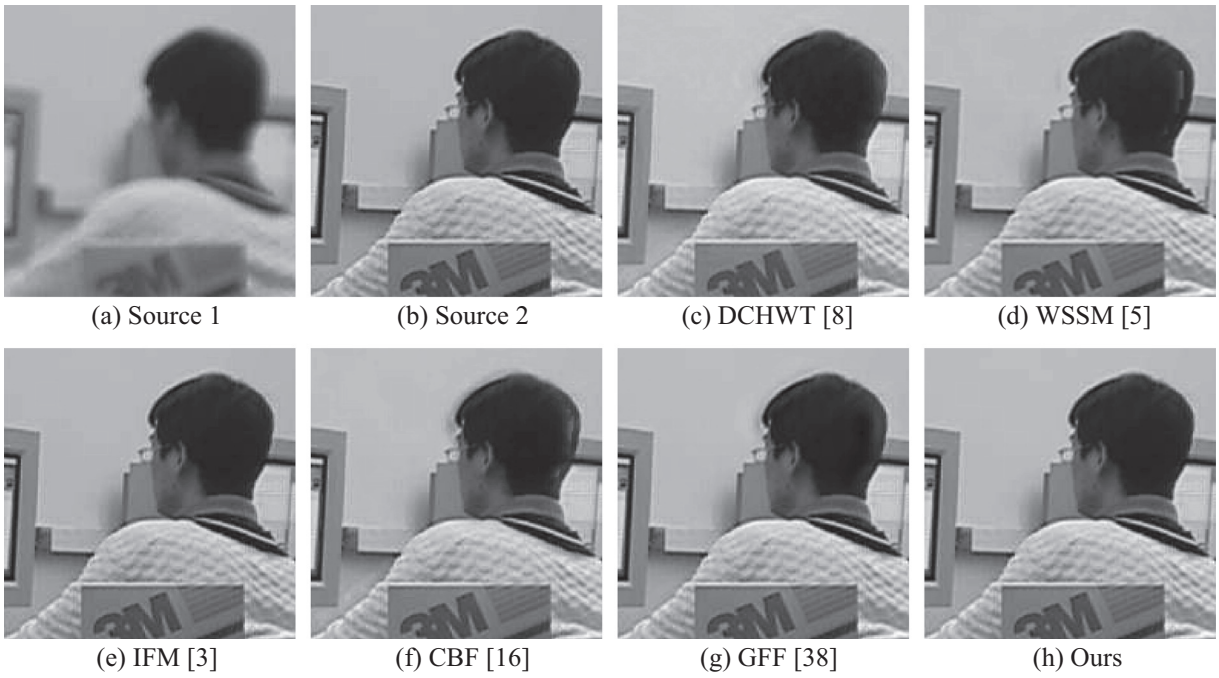


Fig. 10. Magnified regions of the fused images presented in Fig. 9.

$$NMI = 2\left[\frac{MI(I_F, I_A)}{H(I_F) + H(I_A)} + \frac{MI(I_F, I_B)}{H(I_F) + H(I_B)}\right] \tag{15}$$

where $H(I_A)$, $H(I_B)$ and $H(I_F)$ are the marginal entropy of source and fused images. $MI(I_F, I_A)$ and $MI(I_F, I_B)$ represent the mutual information between the fused image I_F and source images I_A, I_B respectively. The assessing index NMI measures how well the original information from source images is transferred to the fused image [38].

(2) Petrovic's metric $Q^{AB/F}$ [40] is a gradient based quality index which evaluates the fusion performance by measuring how well gradient information of the source images conducted to the fused image. It is calculated by:

$$Q^{AB/F} = \frac{\sum_{i,j}(Q^{AF}(i,\ j)w^A(i,\ j) + Q^{BF}(i,j)w^B(i,\ j))}{\sum_{i,j}(w^A(i,\ j) + w^B(i,\ j))} \tag{16}$$

where $Q^{AF}(i,\ j) = Q_g^{AF}(i,\ j)Q_o^{AF}(i,\ j)$. $Q_g^{AF}(i,\ j)$ and $Q_o^{AF}(i,j)$ are the edge strength and orientation preservation values at location (i,j) [40]. Q^{BF} is computed similar to Q^{AF} and $w^A(i,j)$ and $w^B(i,j)$ are the importance weights of $Q^{AF}(i,j)$ and $Q^{BF}(i,j)$, respectively.

(3) The quality index VIFF [41] is a new image fusion quality metric based on visual information fidelity. In calculation of VIFF, the images are decomposed and visual information from the two source-fused pairs is captured using several models including Gaussian scale mixture (GSM) model, the distortion model, and the HVS model. Effective visual information of the fusion is then measured in all blocks in each sub-band. Finally the visual information of all sub-bands is integrated to an overall quality measure. For more details refer to [41]. The larger the values of the three above mentioned fusion quality metrics are, the better the fusion performance will be.

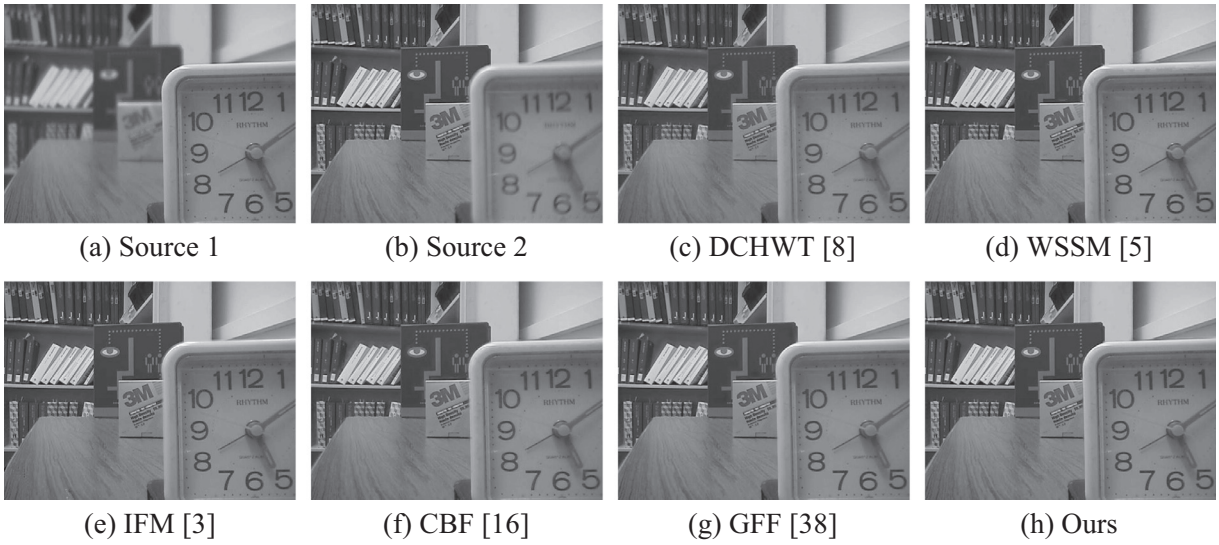


Fig. 11. The "disk" source images and fusion results obtained by different fusion methods.

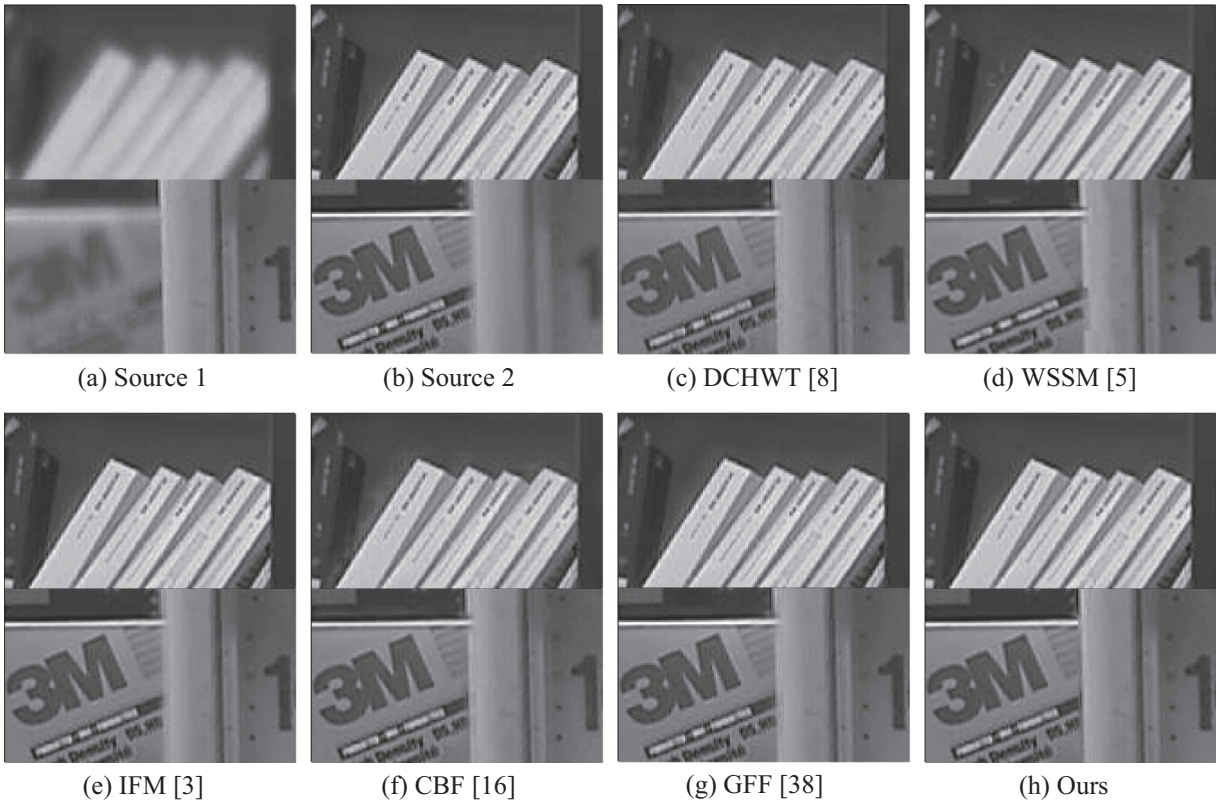


Fig. 12. Magnified regions of the fused images presented in Fig. 11.

4.3. Fusion results and discussions

4.3.1. Evaluation on grayscale dataset

The first experiment is performed on grayscale image dataset. In this dataset, the "book" and "toy" source images which exhibit different level of defocus blur are chosen as the training images. We learn a dictionary of size 400 with sparsity level of $T = 5$ on a collection of 9×9 patches randomly sampled from focus information maps of the training source images. The learned dictionary is presented in Fig. 8(c).

Two examples from fusion results of grayscale multi-focus dataset are shown in Figs. 9 and 11 for visual comparison. Fig. 9(a) and (b) show the "lab" source images and the fused image of them with different fusion methods based on DCHWT, WSSM, IFM, CBF, GFF, and our proposed method are presented in Fig. 9(c)–(h). It can be observed that the DCHWT-based method suffers from ringing effect and it also losses edge contrast to some degree. The fused image obtained by WSSM method contains clear artifacts and artificial edges particularly in flat regions of source images such as the man's hair and the background wall. Results

Table 1
Quantitative assessments of different multi-focus image fusion methods for grayscale dataset. Best results are in bold.

Source images	Quality measure	Method					
		DCHWT	WSSM	CBF	IFM	GFF	Ours
Clock	$Q^{AB/F}$	0.6939	0.7094	0.7258	0.7328	0.7328	**0.7395**
	VIFF	0.9026	0.9421	0.9234	0.9391	0.9391	**0.9443**
	NMI	0.9771	1.1542	1.0969	1.1123	1.1123	**1.2501**
Disk	$Q^{AB/F}$	0.6549	0.6659	0.6991	0.7245	0.7256	**0.7391**
	VIFF	0.8301	0.8695	0.8639	0.8781	**0.8838**	0.8811
	NMI	0.8361	0.8737	0.9203	1.0943	0.9731	**1.1503**
Lab	$Q^{AB/F}$	0.6625	0.6708	0.7121	0.7384	0.7380	**0.7480**
	VIFF	0.8531	0.9005	0.8899	0.9112	0.9150	**0.9176**
	NMI	1.0025	1.0241	1.0690	1.2221	1.1332	**1.2684**
Leaf	$Q^{AB/F}$	0.6876	0.7311	0.7205	0.7459	0.7511	**0.7620**
	VIFF	0.8018	0.9128	0.8562	0.9135	0.9230	**0.9269**
	NMI	0.6288	0.7733	0.7261	0.9054	0.8135	**1.0091**
Pepsi	$Q^{AB/F}$	0.7527	0.7289	0.7689	0.7710	0.7777	**0.7820**
	VIFF	0.8120	**0.8853**	0.8413	0.8788	0.8806	0.8738
	NMI	0.9657	1.0291	1.0188	1.1056	1.0393	**1.2522**
Wafer	$Q^{AB/F}$	0.5692	0.6092	0.6159	0.6333	0.6557	**0.7087**
	VIFF	0.7157	0.7382	0.7546	0.6979	0.7657	**0.7766**
	NMI	0.5296	0.5566	0.5825	0.7275	0.6179	**0.9471**

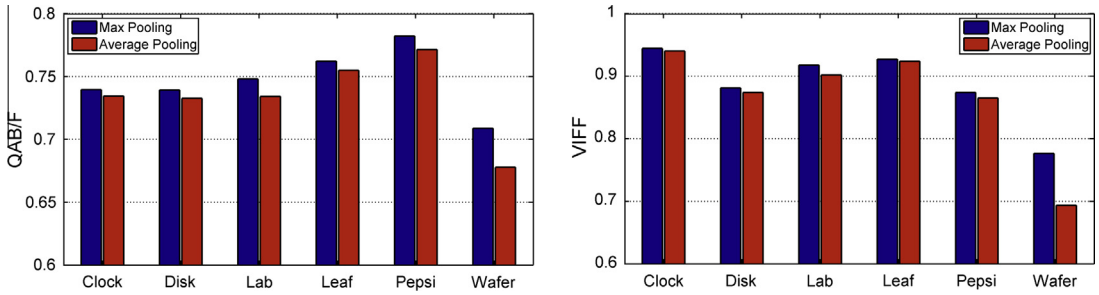


Fig. 13. Performance of our fusion method with different pooling schemes.

of CBF and GFF methods also show artifacts on the head of the man where due to movement is not registered in source images. GFF is more robust to image miss-registration than CBF. We can see significant artifacts around the man's hand in CBF result. In addition, CBF fusion result shows halos near some edges arisen from the bilateral filter used in this method. For clearer comparison, close-up views of the man are presented in Fig. 10. Although the IFM method performs well in this area, it exhibits artifacts on other parts such as the desk region close to the clock. We can see that the fused image produced by our method exhibits the best visual quality with no obvious artifacts.

The fusion results of the "disk" image set from the grayscale multi-focus dataset are shown in Fig. 11. The "disk" source images are shown in Fig. 11(a) and (b) and the fused images based on DCHWT, WSSM, IFM, CBF, GFF, and our proposed method are presented in Fig. 11(c)–(h). Similar to previous example, the fused image obtained by DCHWT method demonstrate severe ringing artifacts which make the entire image blur. The WSSM-based method produces a sharp image but shows serious artifacts around edges, e.g. clock borders. Results of CBF and GFF methods have halo artifacts near edges, but degradation of the CBF fused image is more significant. The IFM method works well in most parts of the fused image however the clock boundary edge is blurred and artifacts can be observed in the left-bottom region of the fused image. Our algorithm performs the best compared to the others in which the focused regions of the source images is preserved without introducing artifacts. For better comparison, close-up views are presented in Fig. 12.

In order to evaluation of fusion performances objectively, the normalize mutual information (NMI), $Q^{AB/F}$, and visual information fidelity (VIFF) are used as image fusion quality measures. The quantitative results of different fusion methods for grayscale multi-focus dataset are shown in Table 1 where training image sets are excluded and the best results are indicated in bold. The parameters of MRF-based regularization in our method are set to $\lambda = 15$ and $\sigma = 5$ which give the best results. It can be seen from Table 1 that the proposed fusion method outperforms the state-of-the-arts in terms of NMI, $Q^{AB/F}$, and VIFF and takes almost all the largest quality indexes.

As mentioned in Section 3.2, max pooling is used in our method to construct the pooled features h_A and h_B from sparse representations of training focus features. We also evaluated the fusion performance of our algorithm when average pooling is used instead of max pooling. The quantitative assessments for both pooling operations are shown in Fig. 13. As can be seen, larger quality metric values are obtained when the pooled features are constructed using max pooling.

4.3.2. Evaluation on Lytro dataset

The Lytro dataset contains 20 color multi-focus image pairs of the same size. With the same setting similar to grayscale dataset, we learn a dictionary from two image pairs and use all other image

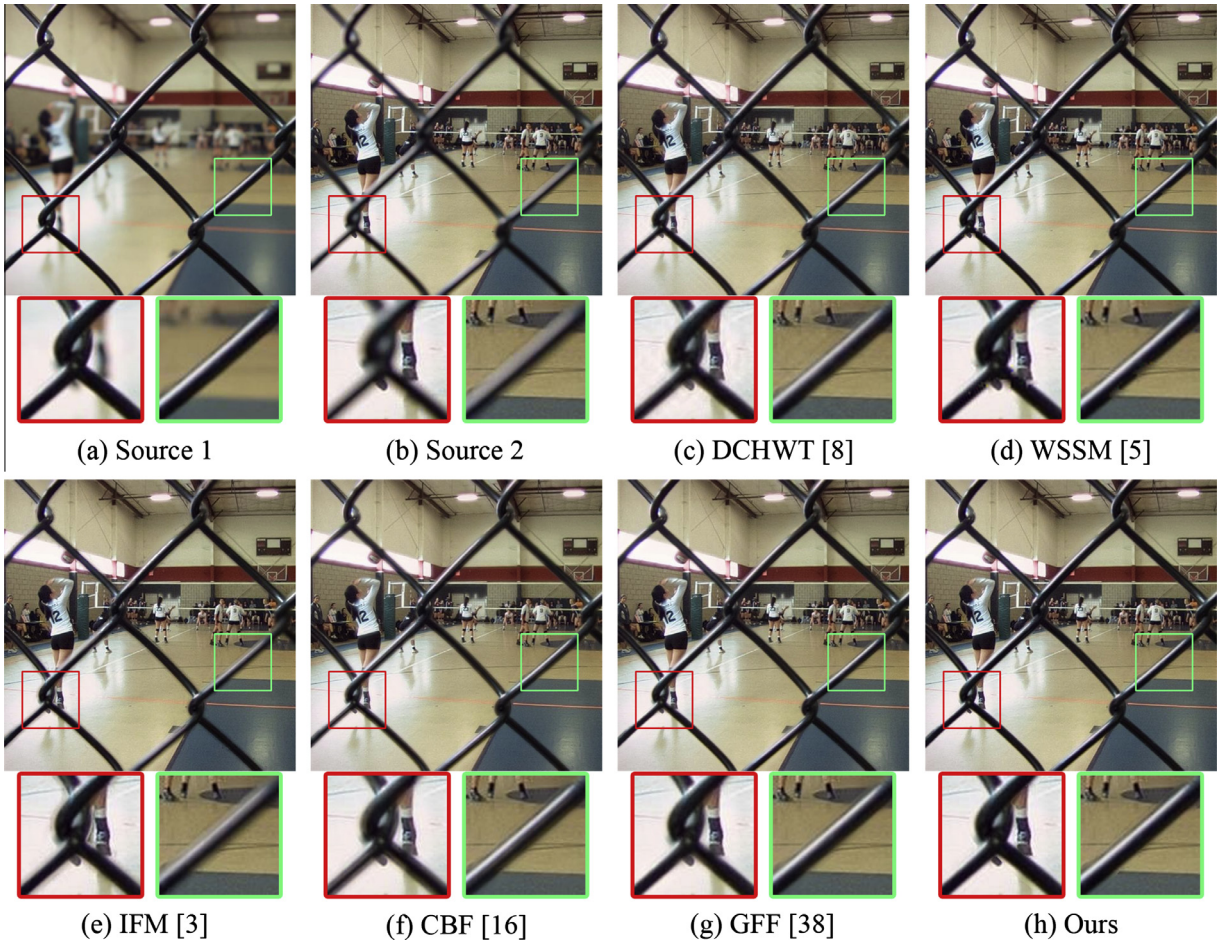


Fig. 14. Color source images from Lytro dataset and fusion results obtained by different methods. Close-up views are shown in the bottom of each result for better visualization. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

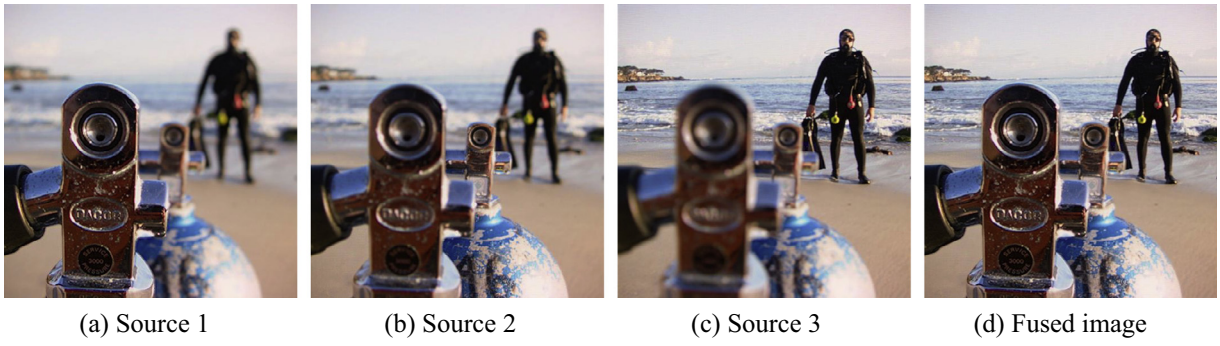


Fig. 15. Application of the proposed method for more than two multi-focus images.

Table 2
Quantitative assessments of different multi-focus image fusion methods for Lytro dataset. Top two results are shown in **Bold** and *italic*.

Quality measure	Method					
	DCHWT	WSSM	CBF	IFM	GFF	Ours
$Q^{AB/F}$	0.7124	0.7296	0.7528	0.7534	*0.7601*	**0.7628**
VIFF	0.8984	0.9319	0.9193	0.9395	*0.9442*	**0.9463**
NMI	0.8971	0.9623	1.0184	*1.1420*	1.0980	**1.1930**

pairs in our experiments. For visual evaluation, the fused images obtained by different methods for a source image pairs of Lytro dataset are demonstrated in Fig. 14. Fig. 14(a) and (b) shows two color multi-focus source images which contain the image of a gym taken through a fence. This is a difficult test because of narrow cross-section of the fence which in the out-of-focus case we can see some structure of the background through semitransparent regions of the blurred fence.

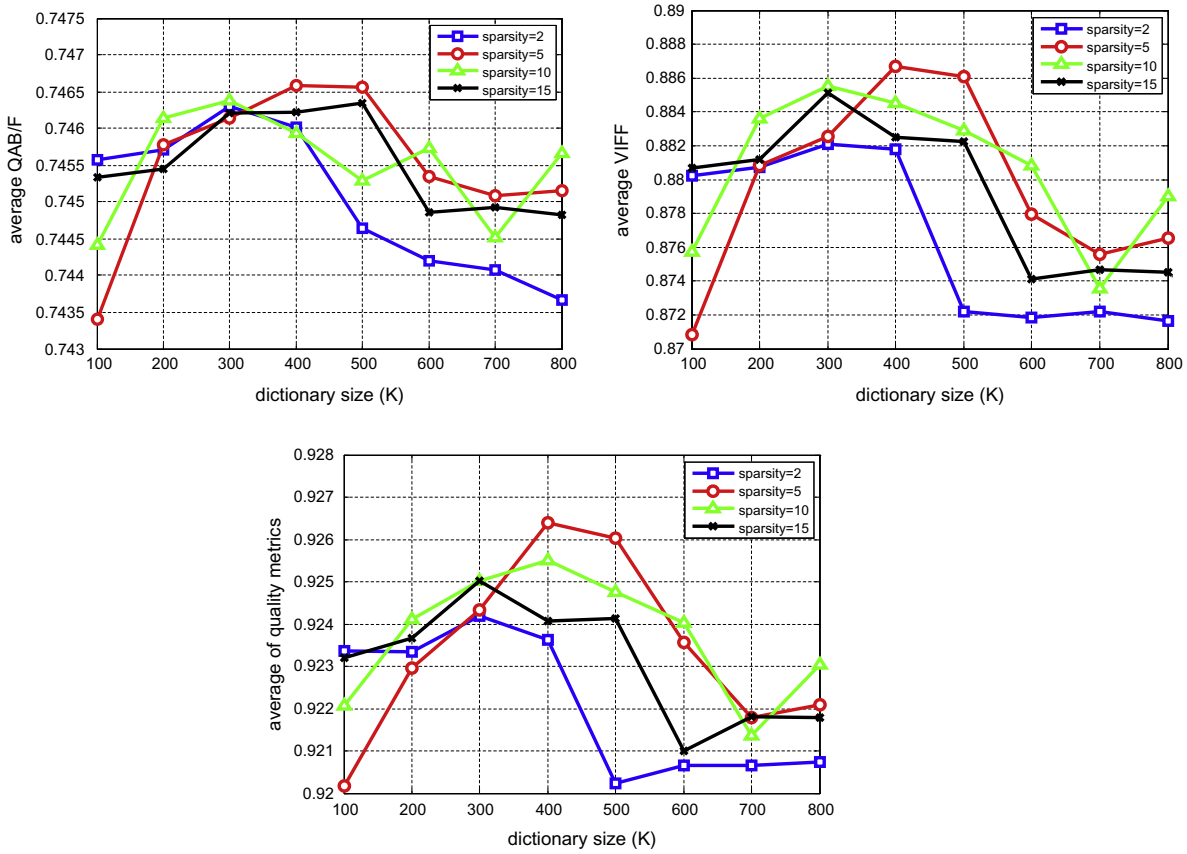


Fig. 16. Average value of fusion quality metrics with respect to the sparsity level and the dictionary size.

Similar to previous experiments on "disk" and "lab" image sets, the DCHWT-based method exhibits undesirable ringing artifacts in the fused image that visually degrade the image. Moreover, the fusion result of this method has lost sharpness especially in the fence areas. As shown in the close-up views of Fig. 14(e) some fence areas in the fused image obtained by the IFM based method are blurred because of erroneous fusing weights on those areas. Severe halo artifacts can be observed around strong edges in the fusion results of CBF. The WSSM fusion method provides a sharp fused image but presents jagged edges in some parts of the fence. However, as shown in Fig. 14(g) and (h), the GFF and the proposed method perform better than the other methods and can well preserve the focused areas of different source images without introducing any artifacts, although our method provides sharper fence in the fused image.

Our method is also applicable when more than two multi-focus images are available. In order to demonstrate it, we also perform the experiment over an image set with more than two source images, shown in Fig. 15(a)–(c). For three source images, the proposed multi-focus fusion method is applied to the two of them at first. Then, the final fusion result is obtained by merging the fused image obtained above with the last source image. As can be seen from Fig. 15(d), all focus regions of source images are integrated into the final result with no obvious artifacts.

At last, the quantitative assessments of different methods for Lytro dataset are shown in Table 2. In that table, the average values of NMI, $Q^{AB/F}$, and VIFF for whole dataset are presented. It can be seen that the proposed method outperforms other ones and gives, in average, the best quality metrics. The GFF method is the second best in terms of $Q^{AB/F}$ and VIFF but the IFM method takes the second place based on the total average of three quality metrics. Overall, it can be concluded based on the experiments, that both by visual comparison and by objective assessments, the proposed method shows competitive fusion performance compared with previous methods.

4.3.3. Dictionary size and sparsity level

We investigated the effects of dictionary size and sparsity level on the fusion performance. In our experiments we tried different settings for these parameters. Intuitively, if the dictionary size is too small, representative elements for sparse coding of focus features will not be sufficient and the pooled features loses discriminant power. If the dictionary size is too large, the pooled features are too sparse and correlation of focus features sparse representation with them may yield zero score values which we cannot make a certain decision in such situations. Fig. 16 shows the average value of quality metrics over all of images of the two datasets when we change the sparsity level along with the dictionary size. As shown in Fig. 16, when the dictionary size is small, fusion performance degrades. On the other hand, the quality measures for dictionary sizes beyond 400 atoms are not improved significantly or decreased while the computational cost increases with the dictionary size. From Fig. 16 it is also observed that for the dictionary size $K = 400$, sparsity factor of 5 achieves better objective performance. Therefore we chose the learned dictionary with $K = 400$ and fixed the sparsity factor $T = 5$ in all the experiments.

5. Conclusions

In this paper, we presented a new multi-focus image fusion method using dictionary-based sparse representation. The proposed method utilized a learned dictionary for sparse coding of focus features extracted from focus information map of multi-focus source images. Correlation of the sparse representations with aggregated sparse coefficients of training samples produced a pixel

level score map which was refined based on guided filter. A spatially smooth and edge-aligned decision map was then obtained using spatial consistency in a MRF optimization framework. Extensive experiments on two multi-focus image datasets were conducted. According to the fusion results, it was observed that the proposed method produced consistent decision maps and was able to well preserve the focused regions of source images with no obvious artifacts. Furthermore, the proposed method achieved competitive fusion performance compared to a number of state-of-the-art fusion methods in terms of visual and objective evaluations. We improved the average quality metrics by 5.4% and 2.4% with respect to the nearest competitor for grayscale and Lytro datasets, respectively.

References

[1] T. Wan, C. Zhu, Z. Qin, Multifocus image fusion based on robust principal component analysis, Pattern Recogn. Lett. 34 (9) (2013) 1001–1008.
[2] S. Li, B. Yang, J. Hu, Performance comparison of different multi-resolution transforms for image fusion, Informat. Fusion 12 (2) (2011) 74–84.
[3] S. Li, X. Kang, J. Hu, B. Yang, Image matting for fusion of multi-focus images in dynamic scenes, Informat. Fusion 14 (2) (2013) 147–162.
[4] Z. Liu, K. Tsukada, K. Hanasaki, Y.K. Ho, Y.P. Dai, Image fusion by using steerable pyramid, Pattern Recogn. Lett. 22 (9) (2001) 929–939.
[5] J. Tian, L. Chen, Adaptive multi-focus image fusion using a wavelet-based statistical sharpness measure, Signal Process. 92 (9) (2012) 2137–2146.
[6] H. Li, B. Manjunath, S. Mitra, Multisensor image fusion using the wavelet transform, Graph. Models Image Process. 57 (3) (1995) 235–245.
[7] J.J. Lewis, R.J. O'Callaghan, S.G. Nikolov, D.R. Bull, N. Canagarajah, Pixel- and region-based image fusion with complex wavelets, Informat. Fusion 8 (2) (2007) 119–130.
[8] B.K. Shreyamsha Kumar, Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform, J., Signal, Image Video Process. 7 (6) (2013) 1125–1143.
[9] Q. Miao, C. Shi, P. Xu, M. Yang, Y. Shi, A novel algorithm of image fusion using shearlets, Opt. Commun. 284 (6) (2011) 1540–1547.
[10] L. Tessens, A. Ledda, A. Pizurica, W. Philips, Extending the depth of field in microscopy through curvelet-based frequency-adaptive image fusion, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2007, pp. I-861–I-864.
[11] Q. Zhang, B.L. Guo, Multi-focus image fusion using the nonsubsampled contourlet transform, Signal Process. 89 (7) (2009) 1334–1346.
[12] B. Yang, S. Li, Pixel-level image fusion with simultaneous orthogonal matching pursuit, Informat. Fusion 13 (1) (2012) 10–19.
[13] B. Yang, S. Li, Multifocus image fusion and restoration with sparse representation, IEEE Trans. Instrum. Meas. 59 (4) (2010) 884–892.
[14] L. Chen, J. Li, C. Chen, Regional multifocus image fusion using sparse representation, Opt. Express 21 (4) (2013) 5182–5197.
[15] Z. Wang, Y. Ma, J. Gu, Multi-focus image fusion using PCNN, Pattern Recogn. 43 (6) (2010) 2003–2016.
[16] B.K. Shreyamsha Kumar, Image fusion based on pixel significance using cross bilateral filter, J. Signal, Image Video Process. (2013) 1–12.
[17] S. Li, J.T. Kwok, Y. Wang, Combination of images with diverse focuses using the spatial frequency, Informat. Fusion 2 (3) (2001) 169–176.
[18] S. Li, B. Yang, Multifocus image fusion using region segmentation and spatial frequency, Image Vis. Comput. 26 (7) (2008) 971–979.
[19] T. Guha, R.K. Ward, Learning sparse representations for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 34 (8) (2012) 1576–1588.
[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] M. Elad, M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries, IEEE Trans. Image Process. 15 (12) (2006) 3736–3745.
[22] Q. Li, H. Zhang, J. Guo, B. Bhanu, L. An, Reference-based scheme combined with K-SVD for scene image categorization, IEEE Signal Process. Lett. 20 (1) (2013) 67–70.
[23] Z. Jiang, Z. Lin, L.S. Davis, Label consistent K-SVD: learning a discriminative dictionary for recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (11) (2013) 2651–2664.
[24] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[25] W. Lu, C. Bai, K. Kpalma, J. Ronsin, Multi-object tracking using sparse representation, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada, May 2013, pp. 2312–2316.
[26] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, S. Yan, Sparse representation for computer vision and pattern recognition, Proc. IEEE 98 (6) (2010) 1031–1044.
[27] M. Elad, M. Figueiredo, Y. Ma, On the role of sparse and redundant representations in image processing, Proc. IEEE 98 (6) (2010) 972–982.
[28] V.M. Patel, R. Chellappa, Sparse representations, compressive sensing and dictionaries for pattern recognition, in: Asian Conference on Pattern Recognition (ACPR), Beijing, China, 2011, pp. 325–329.
[29] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online dictionary learning for sparse coding, in: Proceedings of the 26th International Conference on Machine Learning, New York, USA, June 2009, pp. 689–696.
[30] W. Huang, Z.L. Jing, Evaluation of focus measures in multi-focus image fusion, Pattern Recognit. Lett. 28 (4) (2007) 493–500.
[31] Y. Boureau, F. Bach, Y. LeCun, J. Ponce, Learning mid-level features for recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, June 2010, pp. 2559–2566.
[32] Y. Boureau, Learning Hierarchical Feature Extractors for Image Recognition, New York University Department of Computer Science, 2012.
[33] Y. Boureau, J. Ponce, Y. LeCun, A theoretical analysis of feature pooling in visual recognition, in: Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 2010, pp. 111–118.
[34] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1794–1801.
[35] K. He, J. Sun, X. Tang, Guided image filtering, in: European Conference on Computer Vision, Heraklion, Greece, Sep. 2010, pp. 1–14.
[36] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, IEEE Trans. Pattern Anal. Mach. Intell. 23 (11) (2001) 1222–1239.
[37] <http://mansournejati.ece.iut.ac.ir/content/lytro-multi-focus-dataset>.
[38] S. Li, X. Kang, J. Hu, Image fusion with guided filtering, IEEE Trans. Image Process. 22 (7) (2013) 2864–2875.
[39] M. Hossny, S. Nahavandi, D. Creighton, Comments on 'information measure for performance of image fusion', Electron. Lett. 44 (18) (2008) 1066–1067.
[40] C. Xydeas, V. Petrovic, Objective image fusion performance measure, Electron. Lett. 36 (4) (2000) 308–309.
[41] Y. Han, Y. Cai, Y. Cao, X. Xu, A new image fusion performance metric based on visual information fidelity, Informat. Fusion 14 (2) (2013) 127–135.