Information Fusion 13 (2012) 10–19

Contents lists available at ScienceDirect

Information Fusion

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

Pixel-level image fusion with simultaneous orthogonal matching pursuit

Bin Yang, Shutao Li *

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

ARTICLE INFO

Article history:
Received 22 October 2009
Received in revised form 12 April 2010
Accepted 20 April 2010
Available online 28 April 2010

Keywords:
Multi-sensor fusion
Image fusion
Simultaneous orthogonal matching pursuit
Sparse representation
Multiscale transform

ABSTRACT

Pixel-level image fusion integrates the information from multiple images of one scene to get an informative image which is more suitable for human visual perception or further image-processing. Sparse representation is a new signal representation theory which explores the sparseness of natural signals. Comparing to the traditional multiscale transform coefficients, the sparse representation coefficients can more accurately represent the image information. Thus, this paper proposes a novel image fusion scheme using the signal sparse representation theory. Because image fusion depends on local information of source images, we conduct the sparse representation on overlapping patches instead of the whole image, where a small size of dictionary is needed. In addition, the simultaneous orthogonal matching pursuit technique is introduced to guarantee that different source images are sparsely decomposed into the same subset of dictionary bases, which is the key to image fusion. The proposed method is tested on several categories of images and compared with some popular image fusion methods. The experimental results show that the proposed method can provide superior fused image in terms of several quantitative fusion evaluation indexes.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

For one scene, many images can be simultaneously acquired by various sensors with development of numerous imaging sensors. Those images usually contain complementary information which is dependent on the natural properties of the sensors and the way the images are obtained. Image fusion can effectively extend or enhance information of the scene by combining the images captured by different sensors [1,2]. The fused image can improve edge detection, image segmentation and object recognition in medical imaging, machine vision and military applications. In the past two decades, many techniques and software for image fusion have been developed [3–5]. According to the stage at which the information is combined, image fusion algorithms can be categorized into three levels, namely pixel-level, feature-level, and decision-level [6]. The pixel-level fusion combines the raw source images into a single image. Compared to feature or decision-level fusion, pixel-level fusion can preserve more original information [7]. Feature-level algorithms typically fuse the source images using their various feature properties, such as regions or edges [8]. Thus, this kind of methods is usually robust to noise and misregistration. Decision-level fusion algorithms combine image descriptions directly, for example, in the form of relational graphs [9]. But the decision-level fusion methods are very much application dependent [1]. In this paper, we only focus on the pixel-level image fusion problem.

The goal of pixel-level image fusion is to combine visual information contained in multiple source images into an informative fused image without the introduction of distortion or loss of information. In the past decades, many pixel-level image fusion methods have been proposed. In all of those methods, multiscale transform based methods are the most successful category of techniques. Typical multiscale transforms include the Laplacian pyramid [10], morphological pyramid [11], discrete wavelet transform (DWT) [12–14], gradient pyramid [15], stationary wavelet transform (SWT) [16,17], and dual-tree complex wavelet transform (DTCWT) [18,19]. Recently developed multiscale geometry analysis, such as ridgelet transform [20], curvelet transform (CVT) [21], the nonsubsampled contourlet transform (NSCT) [22,23], are also applied to image fusion. There are three basic steps for multiscale transform based image fusion: firstly, the source images are decomposed into multiscale representations with different resolutions and orientations. Then the multiscale representations are integrated according to some fusion rules. Finally, the fused image is constructed using the inverse transform of the composite multiscale coefficients [6,13].

Multiscale transform based image fusion methods assume that the underlying information is the salient features of the original images, which are linked with the decomposed coefficients [6]. This assumption is reasonable for the transform coefficients that correspond to the transform bases which are designed to represent the important features, such as edges and lines of an image.

* Corresponding author. Tel.: +86 731 88828850; fax: +86 731 88822224.
E-mail addresses: yangbin01420@163.com (B. Yang), shutao_li@yahoo.com.cn (S. Li).

1566-2535/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.inffus.2010.04.001

However, because wavelets and related multiscale transforms are working with a limited dictionary, each multiscale transform has its own advantages and disadvantages. For example, wavelets are suitable to represent image details [24]. Ridgelet bases are suitable to express lines [20], and discrete cosine transform (DCT) bases to textures which are periodic. Then, what is the optimal representation of the underlying information of source images? It is obvious that extracting the underlying information of the original images effectively and completely would make the fused image more accurate [15].

Sparse representation describes natural signals, such as images, by a sparse linear combination of atoms, which are columns of a overcomplete dictionary [25,26]. Sparsity and overcompleteness are two major characteristics of signal sparse representation. Overcompleteness means that the number of atoms exceeds the number of image pixels. It makes the dictionary richer, which usually contains combined transform bases or non-parametric trained dictionaries [27]. And it also allows for more meaningful representations than the traditional wavelet transform method. Thus, common image features can be well described by only 'a few descriptions' of the sparse representation coefficients [28]. Sparsity means that the coefficients are sparse, which leads to efficient algorithms for handling natural signals. Applications that benefit from the sparsity and overcompleteness include compression, regularization in inverse problems [29], image and video compression [30,31], computer vision tasks and so on [32].

In this paper, using the 'a few descriptions' as the salient features of the original images, we propose a sparse representation theory based image fusion method. Based on the conception that image fusion depends on local information of source images, we conduct the sparse representation on overlapping patches instead of the whole image, so that a small size of dictionary is needed. In addition, the proposed image fusion framework requires that different source images are decomposed into the same subset of dictionary bases. So, the simultaneous orthogonal matching pursuit (SOMP) [33] is used to resolve this problem. The sparse representation is conducted with three kinds of typical overcomplete dictionaries, namely the overcomplete DCT bases consisting of 256 prototype signal-atoms, the hybrid dictionary consisting of DCT bases, wavelet 'db1' bases, Gabor bases, and ridgelet bases, and a trained dictionary obtained from learning natural sample data.

The outline of this paper is as follows: in Section 2, we briefly review the signal sparse representation theory. The SOMP based multi-sensor image fusion scheme is proposed in Section 3. Experimental results are presented in Section 4, where the proposed method is compared with some popular ones, especially the methods based on multi-scale transforms. We conclude in Section 5 with some discussions and future work.

2. Signal sparse representation theory

Sparse representation is based on the idea that a signal can be constructed as a linear combination of atoms from a dictionary, where the number of atoms in the dictionary is larger than the signal dimension. This means that the dictionary is overcomplete, so there are numerous ways to represent the signal, among which sparse representation refers to the one with the fewest atoms. It is recognized as the simplest form in all the representations, and is a powerful model in signal processing [26–30]. In this paper, we use bold-face capital letter, $\mathbf{D}$, to denote a matrix. Vectors are written as bold-face lower-case letters. The inner product of two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{R}^n$ is defined as $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{n} \boldsymbol{x}(i)\boldsymbol{y}(i)$. For signals from a class $\Gamma \subset \mathbf{R}^n$, the sparse representation theory assumes that there exists a dictionary $\mathbf{D} \in \mathbf{R}^{n \times T}$, $n < T$, which contains T prototype signals that refer to atoms. For a signal $\boldsymbol{x} \in \Gamma$, the sparse representation finds a very

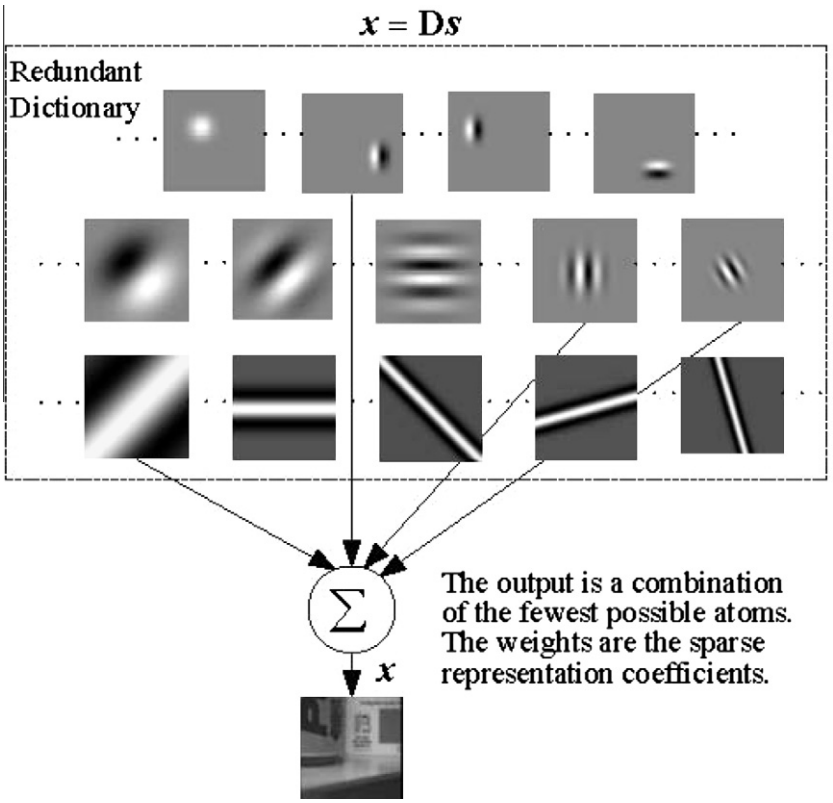

Fig. 1. Sparse representation of an image.

sparse vector $\boldsymbol{s} \in \mathbf{R}^T$, such that $\mathbf{D}\boldsymbol{s} = \boldsymbol{x}$. More formally, the sparse vector can be obtained by solving the following optimization problem:

$$\hat{\boldsymbol{s}} = \arg\min_{\boldsymbol{s}} \|\boldsymbol{s}\|_0 \text{ subject to } \mathbf{D}\boldsymbol{s} = \boldsymbol{x} \quad \text{or} \quad \|\mathbf{D}\boldsymbol{s} - \boldsymbol{x}\|_2^2 < \varepsilon, \tag{1}$$

where $\|\boldsymbol{s}\|_0$ denotes the number of nonzero components in $\boldsymbol{s}$. $\hat{\boldsymbol{s}}$ is the sparse representation of signal $\boldsymbol{x}$ with overcomplete dictionary $\mathbf{D}$. When Γ refers to a class of images, an image can be seen as a superposition of a number of image atoms selected from an overcomplete dictionary $\mathbf{D}$ as depicted in Fig. 1.

The dictionary can be chosen from overcomplete transform bases or be constructed as a hybrid of several multiscale transform bases. It can also be trained to fit a given set of signal examples [27]. The overcomplete transform bases can be the overcomplete wavelets, the steerable wavelet filters, or the Fourier transforms directly. Thus, the construction in this case is simple. The hybrid dictionary can be constructed by integrating the multiscale Gabor functions, wavelets, libraries of windowed cosines with a range of different widths and locations, and multiscale windowed ridgelets, etc. A trained dictionary can be learned by dictionary learning methods [27,34] from a set of training signals.

3. Proposed fusion scheme

3.1. Simultaneous orthogonal matching pursuit

The sparse representation optimization problem in Eq. (1) is an NP-hard problem [27]. Thus, approximate solutions are considered instead. Two of the most frequently discussed approaches are the matching pursuit (MP) [25] and the orthogonal matching pursuit (OMP) algorithms [26]. They are greedy algorithms that select the dictionary atoms sequentially.

For the image fusion problem, multiple source images need to be decomposed simultaneously. However, for the MP and OMP, the decomposed sparse coefficients of different images may correspond to different subset of atoms of the dictionary. This is similar to a decomposition of each input image patch using a different family of wavelet. Thus, the fusion rule will be hard to design. For image fusion, we hope that the different source images are decomposed into the same subset of dictionary atoms. In this paper, the SOMP technique is employed to solve this problem [33]. The SOMP is a variant of OMP and it assumes that different signals can be constructed from the same sparse set of basis atoms, but with different coefficients. In the SOMP algorithm, a fixed dictionary $\mathbf{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_T]$ is used to represent each of the K signals $(\boldsymbol{x}_k)_{k=1}^K$. Each SOMP iteration selects the column index that accounts for the greatest amount of residual energy across all signals. Fig. 2 gives a detailed description of the SOMP.

When the sparse representation theory introduced in Section 2 is used for image fusion, the 2-D images should be converted into vectors. However, the vector sizes are usually too large. For example, a 256×256 image will be converted into a vector in the length of 65,536 and the size of the corresponding dictionary would be far greater. Therefore, directly applying matching pursuit method to the source images is highly ineffective. For the other methods such as iterative shrinkage/thresholding (IST) [35] and *compressive sampling matching pursuit* (CoSaMP) [36] methods, the dimensional is not the problem. However, they can not guarantee that the different source images are decomposed into the same subset of dictionary atoms. Therefore, considering that image fusion depends on local information of the source images, we conduct the sparse representation on the overlapping patches instead of the whole image. It is defined as the "sliding window" technique: starting from the top-left corner of the source images, a sliding window of fixed size moves pixel by pixel in a raster-scan order over the entire image until the bottom-right corner is reached. Each patch is rearranged to form a vector and is normalized to zero-mean by subtracting the mean value of each patch. The subtracted local means are similar to the low frequency coefficient of the wavelet transform coefficients and are stored for the reconstruction.

3.2. Fusion rule

For pixel-level image fusion algorithms, there are two key issues, namely activity-level measurement and coefficient combination. Activity-level measurement identifies the importance of the transform coefficients of the input images, and coefficient combination transfers the useful information into the fused image [6,13,15]. We notice that the activity-level should be designed to easily determine the quality of each source image. Generally speaking, for the multiscale transform based image fusion, the

Input: Dictionary $\mathbf{D} \in R^{n \times T}$, signals $\{\boldsymbol{x}_k\}_{k=1}^K, \boldsymbol{x}_k \in R^n$, threshold ε, an empty matrix Φ.

Output: Sparsity coefficients $\{\boldsymbol{\alpha}_k\}_{k=1}^K, \boldsymbol{\alpha}_j \in R^T$.

SOMP:

1. Initialize the residuals $\boldsymbol{r}_k^{(0)} = \boldsymbol{x}_k$, for $k = 1, 2, ..., K$, set iteration counter $l = 1$.
2. Select the index $\hat{t}_l$ which indicates the next best coefficient atom to simultaneously provide good reconstruction for all signals by solving:

$$\hat{t}_l = \arg \max_{t=1,2,\ldots,T} \sum_{k=1}^{K} \left| \left\langle \boldsymbol{r}_k^{l-1}, \boldsymbol{d}_t \right\rangle \right|. \tag{2}$$

3. Update sets, $\boldsymbol{\Phi}_l = \left[\boldsymbol{\Phi}_{l-1}, \boldsymbol{d}_{\hat{t}_l}\right]$.
4. Compute new coefficients (sparse representations), approximations, and residuals as:

$$\boldsymbol{\alpha}_k^{(l)} = \arg\min_{\alpha} \left\| \boldsymbol{x}_k - \boldsymbol{\Phi}_l \boldsymbol{\alpha} \right\|_2 = \left(\boldsymbol{\Phi}_l^T \boldsymbol{\Phi}_l \right)^{-1} \boldsymbol{\Phi}_l^T \boldsymbol{x}_k, \text{ for } k = 1, 2, ..., K \tag{3}$$

$$\hat{\boldsymbol{x}}_k^{(l)} = \boldsymbol{\Phi}_l \boldsymbol{\alpha}_k^{(l)}, \text{ for } k = 1, 2, ..., K, \tag{4}$$

$$\boldsymbol{r}_k^{(l)} = \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k^{(l)}, \text{ for } k = 1, 2, ..., K. \tag{5}$$

5. Increase the iteration counter $l = l + 1$, If $\sum_{k=1}^{K} \left\| \boldsymbol{r}_k^{(l)} \right\|_2^2 > \varepsilon^2$, go back to 2.

Fig. 2. The SOMP algorithm.

activity-level is described by the absolute value of the corresponding transform coefficients. As we know, the wavelet transform projects the source images onto the localized bases, which are usually designed to represent the sharpness or the edges of an image [37]. Therefore, transformed coefficients (each corresponds to a transform basis) of an image are meaningful to detect and emphasize salient features. The larger the absolute value of the coefficients is, the more information it contains.

From sparse representation theory described in Section 2, each of the sparse coefficients corresponds to an atom. As wavelet bases, the sparse representation dictionary atoms are also designed to represent the sharpness or edges which represent salient features of an image. Thus, the activity-level can be also described by the absolute value of the corresponding coefficient in the sparse representation coefficients.

Averaging and absolute maximum are usually used as the fusion rules [6,13]. The averaging rule means that the corresponding coefficients are averaged by the weight which is depended on the activity-level. It preserves the contrast information of the source images. However, the detailed features such as the edge or the lines would get smoother. For the coefficients combining method, we hope that it could integrate the visual information contained in all the source images into fused image with no distortion or loss of information. However, transforming all the visual information from input source images into the fused image is almost impossible. As a result, a more practical way for image fusion is that the fused image is represented from only the most important input information. Therefore, the choosing absolute maximum (max-abs) rule is usually used for image fusion. The combined coefficients are obtained by selecting entry of coefficients with maximum absolute value. As mentioned before, the local mean of each patch is similar as the low frequency coefficients of the wavelet transform. Thus the local mean of each patch is combined by averaging technique which is often used in multiscale transform based image fusion schemes [6,13].

3.3. Fusion scheme

The image fusion scheme using image sparse representation theory is summarized in Fig. 3. In image fusion, it is essential that the source images are registered, which means that the objects in all images are geometrically aligned. Blanc et al. [38] showed that a small geometrical distortion may produce a noticeable effect on the quality of fused images. In addition, Zhang and Blum discussed the use of image registration in image fusion in [39]. In, fact, the task of registration is very challenging, particularly when images are captured with cameras where extrinsic and/or intrinsic parameters are different. Thus, many researchers study the image registration problem separable with the image fusion, and many effective methods for image registration have been proposed. In this paper, we assumed the K source images, $\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_K$ with size of $M \times N$, are geometrically registered. The whole fusion procedure of the proposed method in this paper takes the following steps:

(1) Divide each source image $\mathbf{I}_j$, from left-top to right-bottom, into every possible image patches of size $\sqrt{n} \times \sqrt{n}$, which is the same as the size of the atom in dictionary. Then all the patches are transformed into vectors via lexicographic ordering and $\{\boldsymbol{v}_k^i\}_{i=1}^{n\times((M-\sqrt{n}+1)\cdot(N-\sqrt{n}+1))}$ is obtained.
(2) Decompose the vectors at each position, i, with different source images, $\{\boldsymbol{v}_k^i\}_{k=1}^K$, into theirs sparse representations, $\boldsymbol{\alpha}_1^i, \boldsymbol{\alpha}_2^i, \ldots, \boldsymbol{\alpha}_K^i$, using the SOMP described in Fig. 2.
(3) Combine the sparse coefficient vectors using the max-abs rule:

$$\boldsymbol{\alpha}_F^i(t) = \boldsymbol{\alpha}_{\hat{k}}^i(t), \quad \hat{k} = \arg\max_{k=1,2,\ldots,K} \left(\left|\boldsymbol{\alpha}_k^i(t)\right|\right), \tag{6}$$

where $\boldsymbol{\alpha}_k^i(t)$ is the tth value of vector $\boldsymbol{\alpha}_k^i$, $t = 1, 2, \ldots, T$. The fused vector $\boldsymbol{v}_F^i$ is obtained by:

$$\boldsymbol{v}_F^i = \mathbf{D}\boldsymbol{\alpha}_F^i. \tag{7}$$

(4) Iterate all the vectors, and reconstruct the fused image, $\mathbf{I}_F$. Firstly, each vector $\boldsymbol{v}_F^i$ is reshaped into a block with size 8×8. Then, the block is added to $\mathbf{I}_F$ at the responding position. Thus, for each pixel position, the pixel value is the sum of several block values. Finally, the pixel value is divided by the adding times at its position to obtain the final reconstructed result $\mathbf{I}_F$.

Notice that two steps in this algorithm make the fusion scheme be shift-invariant, which is of great importance for image fusion. Firstly, the overcompleteness of the dictionary makes the sparse representation is shift-invariant [4]. Secondly the "sliding window" operation is a shift-invariant scheme.

4. Experiments

4.1. Experimental setup

The size of the "sliding window" and the global error of SOMP are two important parameters. As to the size of the "sliding

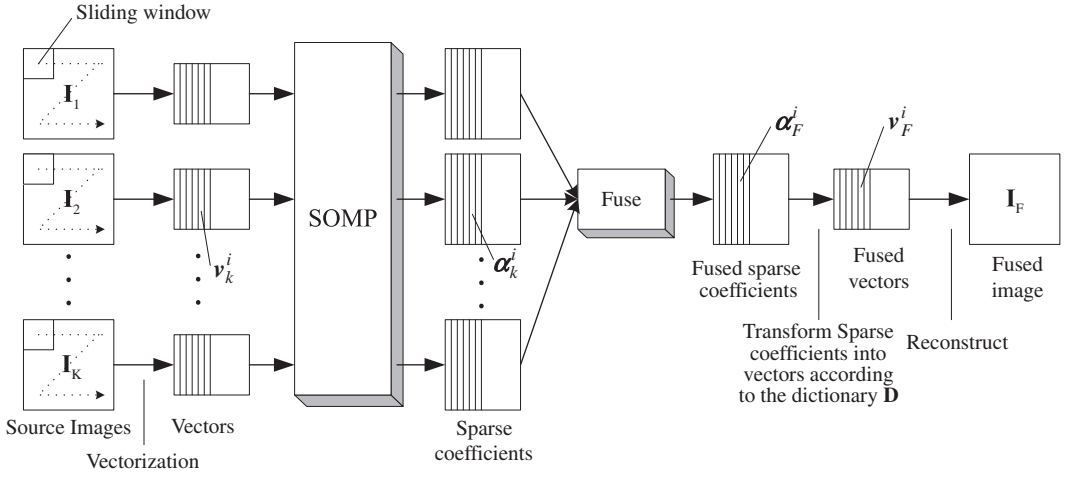

Fig. 3. Procedure of image fusion based on sparse representation.

window", the larger the widow, the bigger the vector transformed from the patch. And the size of the corresponding dictionary also increases according to the sparse theory introduced in Section 2. Thus the process of the SOMP becomes slower. As the size of the "sliding window" decreases, the process of the SOMP becomes faster, but the information contained in the patches would be not sufficient. It may miss some of the important features of the source images. In the following, the patch size is set to 8×8 which has been proved to be appropriate setting for the image denoising application [28]. The second parameter is the global error ε in Eq. (1) which can be tuned according to the noise intensity of the source images. In this paper, we assume that the source images are all clear. So a small global error is set, i.e., $\varepsilon = 0.01$.

Three kinds of overcomplete dictionaries are used to test the performance of the proposed method. The first one is the overcomplete DCT bases. We sample the cosine wave in different frequencies to result 16 vectors with length of 8. The 2-D overcomplete separable version of the DCT bases consist of all possible tensor products of 1-D bases. Then, the overcomplete DCT dictionary are constructed by lexicographically ordering the 2-D overcomplete DCT bases into vectors with length of 256. The 2-D overcomplete separable version of the DCT bases are presented in Fig. 4a. The second one is the hybrid dictionary which consists of DCT bases, wavelet 'db1' bases, Gabor bases, and ridgelet bases, as shown in Fig. 4b. The 1-D 'db1' bases function is

$$\psi_i^j(x) := \psi(2^j x - i) \quad i = 0, \ldots, 2^j - 1, \tag{8}$$

where

$$\psi(x) := \begin{cases} 1 & \text{for } 0 \leqslant x < 1/2 \\ -1 & \text{for } 1/2 \leqslant x < 1 \\ 0 & \text{otherwise} \end{cases}$$

The 2-D wavelet basis consists of all possible tensor products of 1-D basis functions. The Gabor bases function is defined as:

$$G(x, y, k_x, k_y) = \exp\left\{\frac{-(x-X)^2 + (y-Y)^2}{2\sigma^2}\right\} \cdot e^{j(k_x x + k_y y)}, \tag{9}$$

where x and y represent the spatial coordinates while k_x and k_y represent the frequency coordinates. X and Y are the spatial localizations of the Gaussian window. The 2D ridgelet is defined using a wavelet function as:

$$\psi_{a,b,\theta} = a^{-1/2}\psi((x_1 \cos\theta + x_2 \sin\theta - b)/a), \tag{10}$$

where $\psi(\cdot)$ is a wavelet function. Each kind of bases consisted of 64 nonredundant bases, thus the size of the overcomplete dictionary is also 256. The third one is the trained dictionary obtained from learning natural sample using the iterative K-SVD algorithm [27], which has been proved effective when it is used to train the overcomplete dictionary. The training data consisted of 50,000 8×8 patches, randomly takes from a database of 50 natural images. A fraction collection of the 50 natural images and the blocks are presented in Fig. 4d. The iteration is set to 200. The obtained dictionary is shown in Fig. 4c.

We test the proposed method on several pairs of source images including computed tomography (CT) and magnetic resonance imaging (MRI) images, infrared and visual images, and optical multi-focus images, shown in Fig. 5. The images shown in Fig. 5a and c are two CT images that show structures of bone, while the images shown in Fig. 5b and d are two MRI images that show areas of soft tissue. In clinical applications, the combined images showing clearly the position of both bone and tissue can aid in diagnosis of doctors. Fig. 5e–h gives two pair of infrared and visual source images. In the infrared image, the object (the people), in Fig. 5e, is clear, while in visual image the background, such as the tree and the road in Fig. 5f, is clear. Many applications need more comprehensive information containing in both of infrared and visual images. Fig. 5i–l shows two pair of multifocus source images. Fig. 5i is near focused where the small clock is in focus and clear while the larger one is out of focus and blurred. Fig. 5j is far focused, and the situations for the clocks with different sizes are contrary. The fused image should contain both the clear clocks in Fig. 5i and j.

The proposed method is compared with some well-known methods based on multiscale transforms including DWT, DTCWT, SWT, CVT and NSCT. For all of those methods, the most popular setting, the max-abs fusion rule, is selected. And for each method, three levels decomposition are used. For DWT and SWT based methods, the wavelet basis is 'db6'. We note that DWT and CVT are shift-variant transform because of the decimating

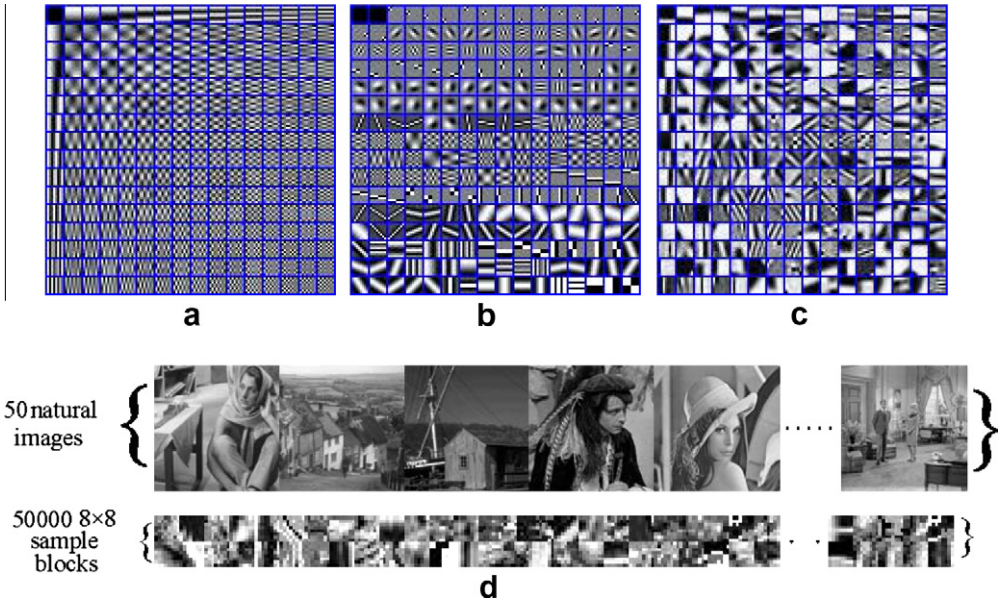

Fig. 4. The overcomplete dictionaries and the training data. (a) The overcomplete DCT dictionary; (b) the hybrid overcomplete dictionary with DCT bases, wavelet 'db1' bases, Gabor bases, and ridgelet bases; (c) the trained overcomplete dictionary; and (d) 50 natural images and the sampled blocks.

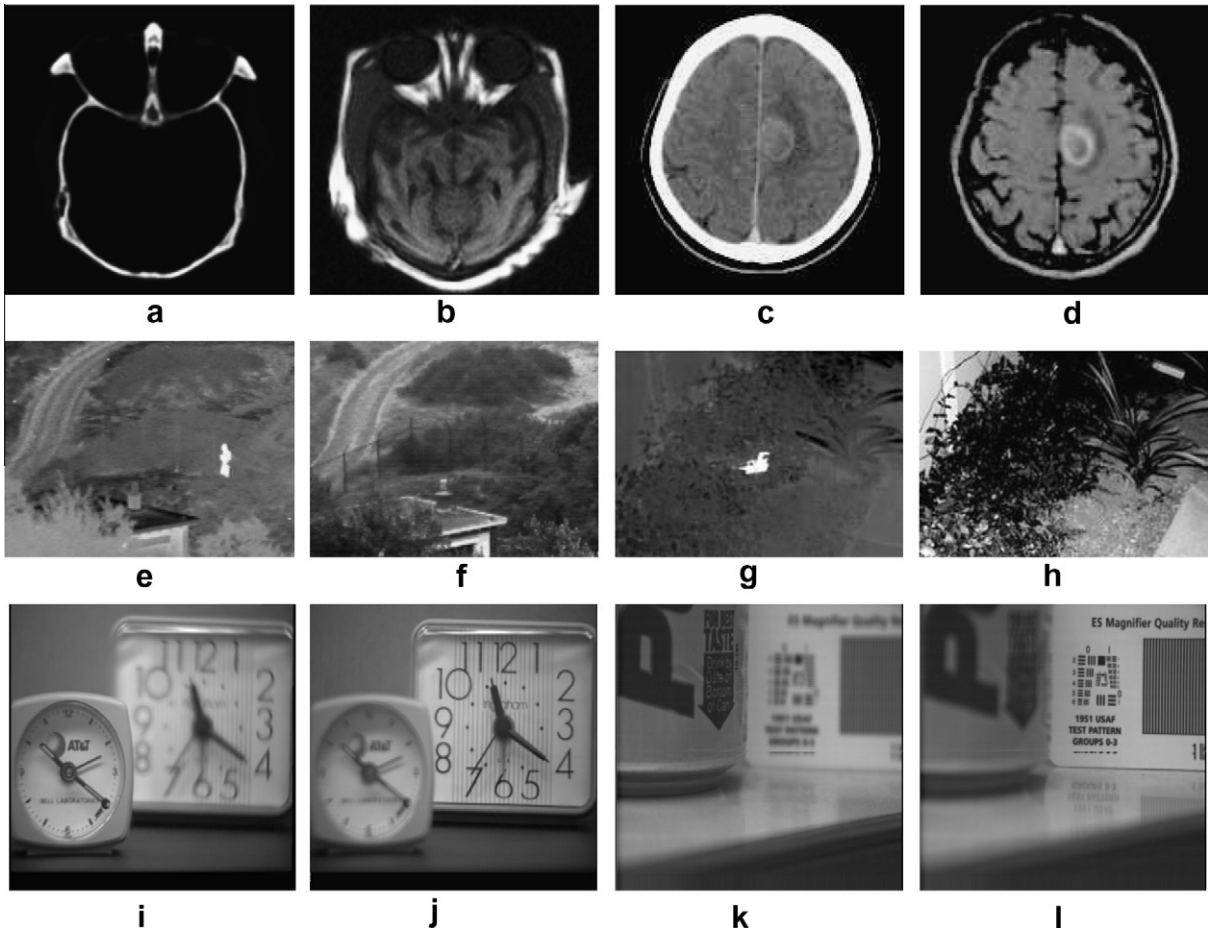

Fig. 5. Source images. The top row: medical images; the middle row: infrared and visual images; the bottom row: multi-focus images.

operator. Thus, the fused images would have Gibbs effect in some degree. So, we combine the common Cycle Spinning algorithm [40] to the DWT and CVT based methods to lessen the Gibbs phenomenon of the fused images. All the experiments are implemented in Matlab 6.5 and on a Pentium(R) 1.7-GHz PC with 512 M RAM.

In this paper, five objective evaluation measures, mutual information (MI) [41], $Q^{AB/F}$ [42], Q_W, Q_E, and Q_0 [43,44] which have been proved to be validated in large degree, are considered to quantitatively evaluate the fusion performances.

(1) MI [41] is a metric defined as the sum of mutual information between each source image and the fused image. Considering the source image A and the fused image F, the mutual information between them is calculated by:

$$I_{AF} = \sum_{a,f} p_{AF}(a,f) \log \frac{p_{AF}(a,f)}{p_A(a)p_F(f)}, \tag{11}$$

where p_{AF} is the jointly normalized histogram of A and F, p_A and p_F are the normalized histogram of A and F, and a and f represent the pixel value of the image A and F, respectively. Let I_{BF} indicate the mutual information between the source image B and the fused image F. Thus, the mutual information between the source images A, B, and the fused image F is defined as:

$$\mathrm{MI} = I_{AF} + I_{BF}. \tag{12}$$

Obviously, larger MI value indicates better fusion result.

(2) The $Q^{AB/F}$ metric was proposed by Xydeas and Petrovic [42], which evaluates the level of the fusion algorithm in transferring input gradient information into the fused image. It is calculated by:

$$Q^{AB/F} = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M}(Q^{AF}(n,m)w^A(n,m) + Q^{BF}(n,m)w^B(n,m))}{\sum_{n=1}^{N}\sum_{m=1}^{M}(w^A(n,m) + w^B(n,m))}, \tag{13}$$

where $Q^{AF}(n,m) = Q_g^{AF}(n,m)Q_\alpha^{AF}(n,m)$; $Q_g^{AF}(n,m)$ and $Q_\alpha^{AF}(n,m)$ are the edge strength and orientation preservation values at location (n,m), respectively; N and M are the size of images, respectively. $Q^{BF}(n,m)$ is similar to $Q^{AF}(n,m)$. $w^A(n,m)$ and $w^B(n,m)$ reflect the importance of $Q^{AF}(n,m)$ and $Q^{BF}(n,m)$, respectively. The dynamic range of $Q^{AB/F}$ is [0 1], and it should be as close to 1 as possible.

(3) Image quality metric based metrics include the local quality index (Q_0), the weighted fusion quality (Q_W) measure, and the edge dependent fusion quality index (Q_E) [43,44], which assess the pixel-level fusion performance objectively. The metric Q_0 between the source image A and the fused image F is defined as follows:

$$Q_0(A,F) = \frac{2\sigma_{af}}{\sigma_a^2 + \sigma_f^2} \cdot \frac{2\bar{a}\bar{f}}{\bar{a}^2 + \bar{f}^2}, \tag{14}$$

where σ_{af} represents the covariance between A and F; σ_a, σ_f denote the standard deviation of A and F; and $\bar{a}$, $\bar{f}$ represent

the mean value of A and F, respectively. $Q_0(A, B, F)$ is the average between $Q_0(A, F)$ and $Q_0(B, F)$, i.e.,

$$Q_0(A,B,F) = (Q_0(A,F) + Q_0(B,F))/2. \quad (15)$$

The metric Q_W between images A, B, and F is defined as follows:

$$Q_W(A,B,F) = \sum_{w \in W} c(w)(\lambda(w)Q_0(A,F|w) + (1-\lambda(w))Q_0(B,F|w)), \quad (16)$$

where $\lambda(w)$ represents the relative salience of A compared to B in the same window w, and $c(w)$ denotes the normalized salience of the window w. The metric Q_E is defined as follows:

$$Q_E(A,B,F) = Q_W(A,B,F) \cdot Q_W(A',B',F')^{\alpha}, \quad (17)$$

where A', B', and F' are the corresponding edge images of A, B, and F, respectively. Parameter α reflects the contribution of the edge images compared to the original images. α is set to 1 in this paper. The larger the Q_W, Q_E and Q_0 values, the better the fused results.

4.2. Experimental results

The fused images of Fig. 5a and b with different fusion methods based on DWT, DTCWT, SWT, CVT, NSCT, and our proposed methods are shown in Figs. 6a–e. Fig. 6f is the fusion result of the proposed method with overcomplete DCT dictionary, denoted by SOMP-DCT. Fig. 6g and h are the fusion results of the proposed method with overcomplete hybrid and trained dictionaries, denoted by SOMP-hybrid and SOMP-trained respectively. Careful inspection of Fig. 6a reveals that the DWT fused image contains significant reconstruction artifacts and it also losses contrast in some degree. We can see that the results of our method exhibit the best visual quality. The important features from both source images are faithfully reserved in the fused image and no reconstruction artifacts are produced. Since the proposed method also has well shift-invariant property. The image contents like tissues are clearly enhanced. Other useful information like brain boundaries and shape are almost perfectly preserved. Moreover, we also conducted the proposed method by replacing the SOMP with the OMP, while the fusion rules and other settings are the same. The fused results are illustrated in Figs. 6i–k respectively. We can see that the proposed SOMP-based method also provides better visual results.

Because of the lack of space, only the fused results by our proposed method are presented in Fig. 7. The left column shows the results of SOMP-DCT, the middle one shows the results of SOMP-hybrid, and the right one shows the results of SOMP-trained. We can see that the features and detailed information are presented well in the result images. For example, the second row of Fig. 7 depicts the fused images of the infrared and visual images for the proposed method with hybrid overcomplete dictionary. It is clear that the fence from the visual image is well transferred into the fused images. In addition, the details of the tree in the visual image

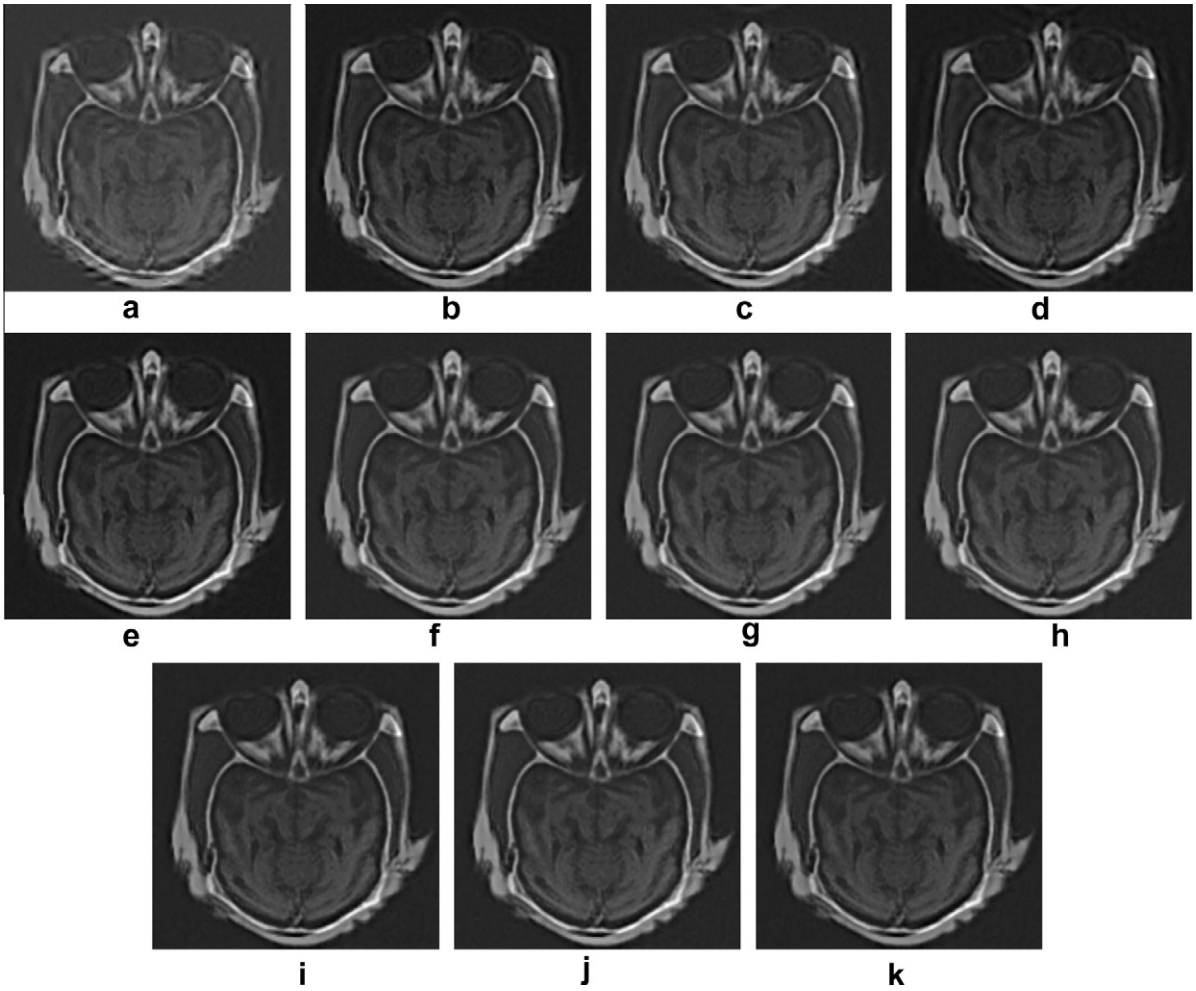

Fig. 6. The fused images: (a) DWT; (b) DTCWT; (c) SWT; (d) CVT; (e) NSCT; (f) SOMP-DCT; (g) SOMP-hybrid; (h) SOMP-trained; (i) OMP-DCT; (j) OMP-hybrid; and (k) OMP-trained.

are visually pleasing and the human figure is much bright in the fused images. The forth row of Fig. 7 depicts the fused images of the multi-focus images for the proposed method. The fused images contain both the clear clocks as shown in the source images in Fig. 5i and j.

The objective evaluations on the fused results of the proposed method and other comparable approaches for the medical images are listed in Table 1. The best results are indicated in bold. We can see from Table 1 that the proposed method takes almost all the largest objective evaluations, which is obviously better than the other methods. The SOMP-trained method provides the best performances for medical image fusion. For the fused results with different dictionaries, the differences between them are not obvious, this may be due to that most of information containing in the source images can be represented by all of the three kinds of dictionaries because of their high redundancy.

Similarly, the objective evaluations of various fusion methods for the infrared and visual images are listed in Table 2. From Table 2 we can see that the proposed method performs best compared with other methods. The SOMP-hybrid method provides the best performances among different dictionaries. The objective evaluations on the fused multi-focus images of various methods are listed in Table 3. The proposed method only loses the Q_E for Fig. 5k and l. The SOMP-DCT method with overcomplete DCT dictionary provides better results than the methods with other two dictionaries. From all of the objective evaluations, we can see that the proposed

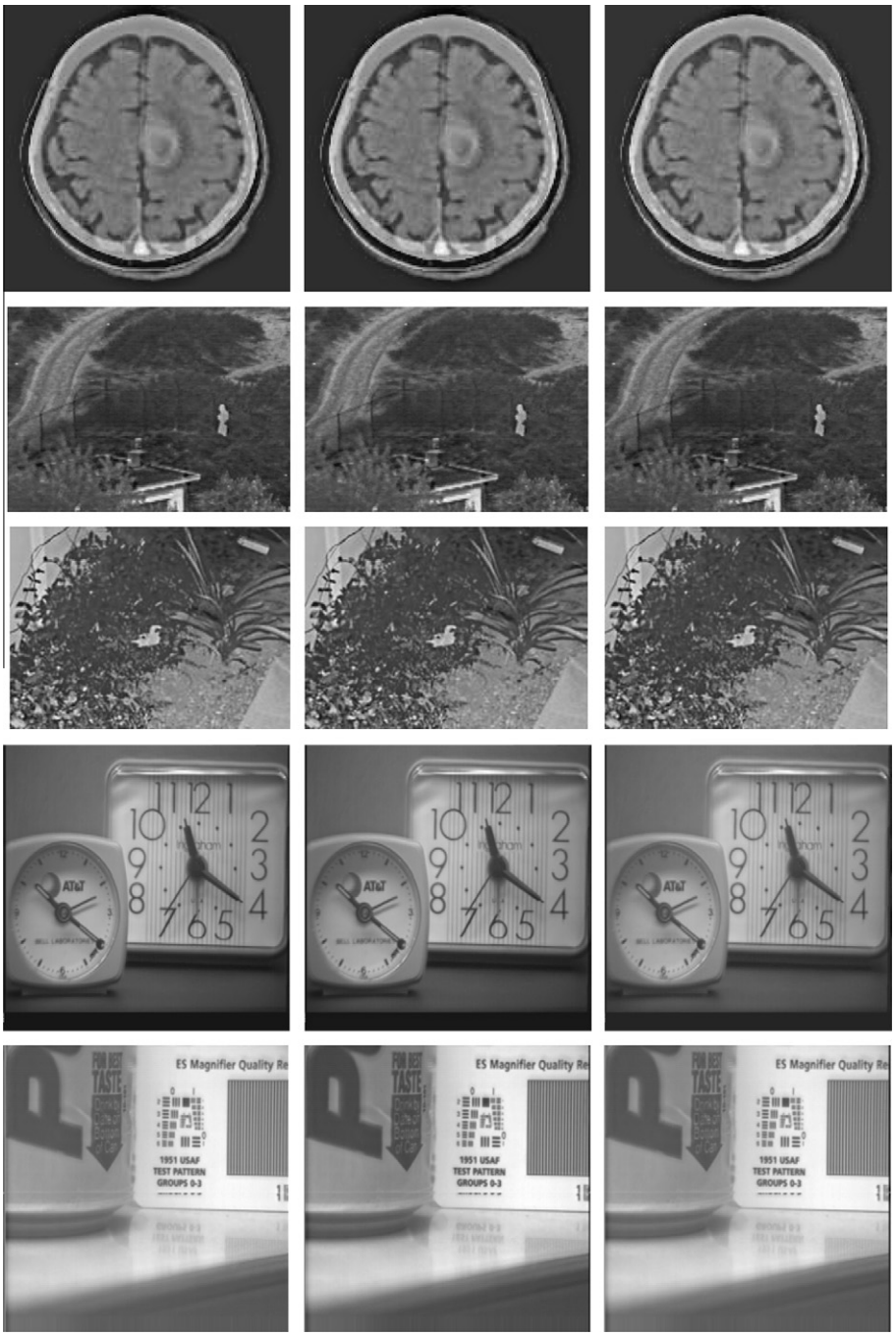

Fig. 7. Fused results of the proposed method. For each row, the left images are from the SOMP-DCT method, the middle ones are from the SOMP-hybrid method and the right ones are from the SOMP-trained based method.

Table 1
Quantitative assessment of various fusion methods for medical images.

Fusion methods	Source images									
	Fig. 5a and b					Fig. 5c and d				
	MI	$Q^{AB/F}$	Q_W	Q_E	Q_0	MI	$Q^{AB/F}$	Q_W	Q_E	Q_0
DWT	2.235	0.600	0.587	0.364	0.555	2.544	0.488	0.567	0.302	0.396
DTCWT	2.250	0.649	0.620	0.411	0.581	2.506	0.521	0.591	0.332	0.431
SWT	2.253	0.650	0.620	0.409	0.586	2.506	0.530	0.586	0.329	0.420
CVT	2.225	0.550	0.516	0.278	0.469	2.521	0.393	0.385	0.136	0.288
NSCT	2.267	0.684	0.641	0.437	0.616	2.517	0.558	0.607	0.353	**0.438**
OMP-DCT	2.293	0.723	0.672	0.470	0.657	2.546	0.571	0.589	0.337	0.397
SOMP-DCT	2.295	0.720	0.680	0.481	**0.660**	**2.561**	0.575	0.617	0.377	0.416
OMP-hybrid	2.294	0.724	0.679	0.477	0.656	2.545	0.578	0.604	0.353	0.406
SOMP-hybrid	2.295	0.727	0.681	0.484	**0.660**	2.557	0.575	0.617	0.374	0.418
OMP-trained	**2.296**	0.722	0.662	0.460	0.651	2.544	0.570	0.579	0.326	0.391
SOMP-trained	**2.296**	**0.737**	**0.682**	**0.487**	**0.660**	2.559	**0.580**	**0.621**	**0.386**	0.417

Table 2
Quantitative assessments of various fusion methods for infrared and visual images.

Fusion methods	Source images									
	Fig. 5e and f					Fig. 5g and h				
	MI	$Q^{AB/F}$	Q_W	Q_E	Q_0	MI	$Q^{AB/F}$	Q_W	Q_E	Q_0
DWT	2.045	0.407	0.613	0.429	0.563	2.334	0.651	0.848	0.723	0.709
DTCWT	2.049	0.437	0.640	0.465	0.589	2.340	0.687	0.861	0.741	0.739
SWT	2.046	0.418	0.635	0.445	0.573	2.337	0.665	0.853	0.737	0.707
CVT	2.045	0.407	0.651	0.455	0.569	2.335	0.654	0.734	0.591	0.588
NSCT	2.050	0.444	0.659	0.484	0.614	2.342	0.694	0.867	0.748	0.749
OMP-DCT	2.048	0.434	0.675	0.485	**0.623**	2.344	0.674	0.873	0.749	0.753
SOMP-DCT	**2.051**	0.462	0.689	0.505	0.636	2.349	0.709	**0.881**	**0.766**	0.764
OMP-hybrid	2.049	0.442	0.683	0.496	0.631	2.346	0.682	0.872	0.746	0.755
SOMP-hybrid	2.050	0.459	**0.691**	**0.509**	**0.637**	**2.350**	**0.714**	**0.881**	**0.766**	0.764
OMP-trained	2.047	0.427	0.659	0.467	0.613	2.342	0.667	0.869	0.741	0.747
SOMP-trained	**2.051**	0.**471**	0.687	0.506	0.635	**2.350**	0.709	**0.881**	**0.766**	**0.765**

Table 3
Quantitative assessments of various fusion methods for multi-focus images.

Fusion methods	Source images									
	Fig. 5i and j					Fig. 5k and l				
	MI	$Q^{AB/F}$	Q_W	Q_E	Q_0	MI	$Q^{AB/F}$	Q_W	Q_E	Q_0
DWT	2.605	0.686	0.866	0.704	0.782	2.663	0.736	0.912	0.825	0.826
DTCWT	2.627	0.698	0.887	0.716	0.795	2.642	0.738	0.911	0.826	0.828
SWT	2.624	0.699	0.876	0.706	0.787	2.656	0.746	0.912	0.826	0.841
CVT	2.611	0.706	0.835	0.653	0.744	2.632	0.741	0.910	0.821	0.812
NSCT	2.643	0.710	0.893	0.727	0.814	2.673	0.753	0.914	**0.829**	0.858
OMP-DCT	2.573	0.673	0.846	0.673	0.764	2.619	0.721	0.886	0.776	0.811
SOMP-DCT	**2.663**	0.712	**0.904**	0.751	**0.842**	**2.681**	**0.762**	0.915	0.825	**0.878**
OMP-hybrid	2.616	0.69	0.884	0.723	0.809	2.657	0.737	0.902	0.804	0.842
SOMP-hybrid	2.661	0.712	0.903	0.744	0.830	2.674	0.760	**0.916**	0.827	0.877
OMP-trained	2.533	0.659	0.818	0.632	0.734	2.567	0.710	0.861	0.746	0.769
SOMP-trained	2.656	**0.714**	0.903	**0.752**	0.840	2.680	0.761	**0.916**	0.817	0.865

method provides higher evaluation index values than the conventional image fusion methods on the whole. And we also see that for different type of source images, the fused results with different overcomplete dictionary have a few difference. Overall, the overcompleted DCT dictionary performs better in fusing the multi-focus images while the hybrid dictionary performs better in fusing the infrared and visual images, and the trained dictionary is more suitable for the medical images. In addition, from the Tables 1–3, we can see that the proposed method also provides very promising results comparing to other methods when replace the SOMP with OMP. However, the performance of the method based on the OMP is obviously lower than that based on the SOMP.

5. Conclusion

This paper presents a novel multi-sensor image fusion algorithm based on signal sparse representation theory. The fusion process is conducted by the simultaneous orthogonal matching pursuit (SOMP). Various multi-sensor images are used to test the

performance of the proposed scheme. The experiments demonstrate that the proposed method provides superior fused image in terms of the pertained quantitative fusion evaluation indexes. In addition, tuning the reconstruction error parameter of the sparse representation according to the noise standard deviation, the proposed fusion scheme is easy to extend to combining of image fusion and restoration problem when the source images are corrupted by noise in acquisition or transmission. However, because the SOMP process is a greedy matching pursuit method, and the "sliding window" scheme is also time-consuming, the computation load of the proposed scheme is heavier than traditional multiscale transform methods. It takes about 30 min to fuse a pair of 256×256 pixel images on our PC. But we notice that the SOMP can be done independently on each pair of patches. So it is easy to parallel implement on any number of multicore processors or FPGA, leading to a substantial speedup.

Acknowledgements

The authors would like to thank the anonymous reviewers for their detailed review, valuable comments, and constructive suggestions. This paper is supported by the National Natural Science Foundation of China (No. 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), and the Open Projects Program of National Laboratory of Pattern Recognition, China.

References

[1] A.A. Goshtasby, S. Nikolov, Image fusion: advances in the state of the art, Information Fusion 8 (2) (2007) 114–118.
[2] V. Petrovic, T. Cootes, Objectively adaptive image fusion, Information Fusion 8 (2) (2007) 168–176.
[3] 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.
[4] S. Daneshvar, H. Ghassemian, MRI and PET image fusion by combining IHS and retina-inspired models, Information Fusion 11 (2) (2010) 114–123.
[5] M. Zribi, Non-parametric and region-based image fusion with bootstrap sampling, Information Fusion 11 (2) (2010) 85–94.
[6] Z. Zhang, R.S. Blum, A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application, Proceedings of the IEEE 87 (8) (1999) 1315–1326.
[7] N. Mitianoudis, T. Stathaki, Pixel-based and region-based image fusion schemes using ICA bases, Information Fusion 8 (2) (2007) 131–142.
[8] G. Piella, A general framework for multiresolution image fusion: from pixels to regions, Information Fusion 4 (4) (2003) 259–280.
[9] M.L. Williams, R.C. Wilson, E.R. Hancock, Deterministic search for relational graph matching, Pattern Recognition 32 (7) (1999) 1255–1516.
[10] P.T. Burt, E.H. Andelson, The Laplacian pyramid as a compact image code, IEEE Transactions on Communications 31 (4) (1983) 532–540.
[11] A. Toet, A morphological pyramidal image decomposition, Pattern Recognition Letters 9 (3) (1989) 255–261.
[12] H. Li, B. Manjunath, S. Mitra, Multisensor image fusion using the wavelet transform, Graphical Models and Image Processing 57 (3) (1995) 235–245.
[13] G. Pajares, J. Cruz, A wavelet-based image fusion tutorial, Pattern Recognition 37 (9) (2004) 1855–1872.
[14] B. Yang, Z.L. Jing, Image fusion using a low-redundancy and nearly shift-invariant discrete wavelet frame, Optics Engineering 46 (10) (2007) 107002.
[15] V.S. Petrovic, C.S. Xydeas, Gradient-based multiresolution image fusion, IEEE Transactions on Image Processing 13 (2) (2004) 228–237.
[16] M. Beaulieu, S. Foucher, L. Gagnon, Multi-spectral image resolution refinement using stationary wavelet transform, in: Proceedings of the International Geoscience and Remote Sensing Symposium, 1989, pp. 4032–4034.
[17] S.T. Li, J.T. Kwok, Y.N. Wang, Discrete wavelet frame transform method to merge Landsat TM and SPOT panchromatic images, Information Fusion 3 (1) (2002) 17–23.
[18] J.J. Lewis, R.J. Ocallaghan, S.G. Nikolov, Pixel- and region-based image fusion with complex wavelets, Information Fusion 8 (2) (2007) 119–130.
[19] J.J. Lewis, R.J. O'Callaghan, S.G. Nikolov, D.R. Bull, C.N. Canagarajah, Region-based image fusion using complex wavelets, in: Proceedings of the 7th International Conference on Image Fusion, 2004, pp. 555–562.
[20] T. Chen, J.P. Zhang, Y. Zhang, Remote sensing image fusion based on ridgelet transform, in: Proceedings of International Conference on Geoscience and Remote Sensing Symposium, 2005, pp. 1150–1153.
[21] 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.
[22] L.D. Cunha, J.P. Zhou, The nonsubsampled contourlet transform: theory, design, and applications, IEEE Transactions on Image Processing 15 (10) (2006) 3089–3101.
[23] B. Yang, S.T. Li, F.M. Sun, Image fusion using nonsubsampled contourlet transform, in: Proceedings on the Fourth International Conference on Image and Graphics, 2007, pp. 719–724.
[24] J.L. Starck, D.L. Donoho, E.J. Candès, Very high quality image restoration by combining wavelets and curvelets, in: Proceedings of SPIE, vol. 4478, 2001, pp. 9–19.
[25] S.G. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries, IEEE Transactions on Signal Processing 41 (12) (1993) 3397–3415.
[26] A.M. Bruckstein, D.L. Donoho, M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, SIAM Review 51 (1) (2007) 34–81.
[27] M. Aharon, M. Elad, A. Bruckstein, The K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation, IEEE Transactions on Signal Processing 54 (11) (2006) 4311–4322.
[28] M. Elad, M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries, IEEE Transactions on Image Processing 15 (12) (2006) 3736–3745.
[29] J. Mairal, M. Sapiro, G. Sapiro, Sparse representation for color image restoration, IEEE Transactions on Image Processing 17 (1) (2008) 53–68.
[30] O. Bryt, M. Elad, Compression of facial images using the K-SVD algorithm, Journal of Visual Communication and Image Representation 19 (4) (2008) 270–282.
[31] A. Rahmoune, P. Vandergheynst, P. Frossard, Sparse approximation using m-term pursuits with applications to image and video compression, Signal Processing Institute, Swiss Federal Institute of Technology EPFL, Tech. Rep. ITS-2005.03, 2005.
[32] J. Mairal, F. Bach, J. Ponce, G. Sapiro, A. Zisserman, Discriminative learned dictionaries for local image analysis, in: Proceeding of the International Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[33] J.A. Tropp, A.C. Gilbert, M.J. Strauss, Algorithms for simultaneous sparse approximation. Part I: greedy pursuit, Signal Processing 86 (3) (2006) 572–588.
[34] K. Engan, K. Skretting, J.H. Husoy, Family of iterative LS-based dictionary learning algorithms, ILS-DLA, for sparse signal representation, Digital Signal Processing 17 (1) (2007) 32–49.
[35] M. Figueiredo, R. Nowak, An EM algorithm for wavelet-based image restoration, IEEE Transactions on Image Processing 12 (8) (2003) 906–916.
[36] D. Needell, J. Tropp, CoSaMP: iterative signal recovery from incomplete and inaccurate samples, Applied and Computational Harmonic Analysis 26 (3) (2009) 301–321.
[37] A. Mahmood, P.M. Tudor, W. Oxford, et al., Applied multi-dimensional fusion, The Computer Journal 50 (6) (2007) 646–659.
[38] P. Blanc, L. Wald, T. Ranchin, Importance and effect of coregistration quality in an example of 'pixel to pixel' fusion process, in: Proceedings of the Second International Conference on Fusion Earth Data, 1998, pp. 67–74.
[39] Z. Zhang, R.S. Blum, A hybrid image registration technique for a digital camera image fusion application, Information Fusion 2 (2) (2001) 135–149.
[40] R.R. Coifman, D.L. Donoho, Translation-invariant de-noising, Wavelets and Statistics, Springer Lecture Notes in Statistics 103 (1995) 125–150.
[41] G.H. Qu, D.L. Zhang, P.F. Yan, Information measure for performance of image fusion, Electronics Letters 38 (7) (2002) 313–315.
[42] C.S. Xydeas, V. Petrovic, Objective image fusion performance measure, Electronics Letters 36 (4) (2000) 308–309.
[43] G. Piella, H. Heijmans, A new quality metric for image fusion, in: Proceedings of the International Conference on Image Processing, 2003, pp. 173–176.
[44] V. Petrovic, Subjective tests for image fusion evaluation and objective metric validation, Information Fusion 8 (2) (2007) 208–216.