Information Fusion 40 (2018) 57–75

Contents lists available at ScienceDirect

Information Fusion

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

Full Length Article

Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review

Qiang Zhang[a,b], Yi Liu[b], Rick S. Blum[c], Jungong Han[d,*], Dacheng Tao[e]

[a] *Key Laboratory of Electronic Equipment Structure Design, Ministry of Education, Xidian University, Xi'an Shaanxi 710071, PR China*
[b] *Center for Complex Systems, School of Mechano-Electronic Engineering, Xidian University, Xi'an Shaanxi 710071, PR China*
[c] *Electrical and Computer Engineering Department, Lehigh University, Bethlehem, PA 18015, United States*
[d] *School of Computing and Communications, Lancaster University, Lancaster, LA1 4WA, UK*
[e] *School of Information Technologies in the Faculty of Engineering and Information Technologies at the University of Sydney, J12/318 Cleveland St, Darlington NSW 2008, Australia*

ARTICLE INFO

Article history:
Received 24 February 2017
Revised 15 May 2017
Accepted 28 May 2017
Available online 9 June 2017

Keywords:
Image fusion
Sparse representation
Dictionary learning
Activity level

ABSTRACT

As a result of several successful applications in computer vision and image processing, sparse representation (SR) has attracted significant attention in multi-sensor image fusion. Unlike the traditional multiscale transforms (MSTs) that presume the basis functions, SR learns an over-complete dictionary from a set of training images for image fusion, and it achieves more stable and meaningful representations of the source images. By doing so, the SR-based fusion methods generally outperform the traditional MST image fusion methods in both subjective and objective tests. In addition, they are less susceptible to mis-registration among the source images, thus facilitating the practical applications. This survey paper proposes a systematic review of the SR-based multi-sensor image fusion literature, highlighting the pros and cons of each category of approaches. Specifically, we start by performing a theoretical investigation of the entire system from three key algorithmic aspects, (1) sparse representation models; (2) dictionary learning methods; and (3) activity levels and fusion rules. Subsequently, we show how the existing works address these scientific problems and design the appropriate fusion rules for each application such as multi-focus image fusion and multi-modality (e.g., infrared and visible) image fusion. At last, we carry out some experiments to evaluate the impact of these three algorithmic components on the fusion performance when dealing with different applications. This article is expected to serve as a tutorial and source of reference for researchers preparing to enter the field or who desire to employ the sparse representation theory in other fields.

1. Introduction

Due to recent technological advancements, extensive varieties of imaging sensors have been employed in many applications including remote sensing, medical imaging, video surveillance, machine vision and security. Thus, finding a way to most effectively utilize the information captured from these multiple sensors, possibly of different modalities, is of considerable interest. Image fusion provides one versatile solution, where multiple aligned images acquired by different sensors are merged into a composite image. The properly fused image is more informative than any of the individual input images and can thus better interpret the scene [1]. As a result, multi-sensor image fusion has always been an active research topic, facilitating a variety of vision-related applications.

To date, a large number of image fusion algorithms have been proposed [2–6], in which multiscale transform-based (MST) fusion methods are the most popular [7–9]. Traditional MST fusion methods are generally those using pyramids [10] and wavelet transforms [11]. Recently developed fusion methods can be considered as their variations and extensions employing multiscale geometric analysis (MGA) tools, such as the Curvelet Transform [12], the Shearlet Transform [13] and the nonsubsampled Contourlet Transform (NSCT) [14]. Thorough reviews on such methods can be found in [2,7].

Sparse representation (SR) [15] has recently drawn significant interest in computer vision and image processing due to its enhanced performance in many applications, such as face recognition [15], action recognition [16], and object tracking [17]. The main idea of SR theory lies in the fact that an image signal can be represented as a linear combination of the fewest possible atoms or transform basis primitives in an over-complete dictionary. Spar-

[*] Corresponding author.
E-mail address: jungonghan77@gmail.com (J. Han).

http://dx.doi.org/10.1016/j.inffus.2017.05.006
1566-2535/© 2017 Elsevier B.V. All rights reserved.

sity means that only a small number of atoms are required to accurately reconstruct a signal, i.e., the coefficients become sparse. Over-completeness indicates that the number of atoms in the dictionary is larger than the dimension of the signal. Thus, a sufficient number of atoms in an over-complete dictionary permit an accurate sparse representation of signals [18].

Not surprisingly, SR has also attracted significant attention in the research field of image fusion [18–21]. Similar to the traditional MST image fusion methods, most of the SR-based image fusion methods also belong to the transform-domain-based techniques.[1] However, there are two main differences between the SR-based and the traditional MST-based fusion methods [18,19].

1. The traditional MSTs usually fix their basis functions in advance for image analysis and fusion. Due to the limitations of predefined basis functions, some significant features (e.g., edges) of source images may not be well expressed and extracted, thereby dramatically degrading the performance of fusion. In contrast, SR generally *learns* an over-complete dictionary from a set of training images for image fusion, which captures intrinsic data-driven image representations tending to be domain agnostic. The over-complete dictionary contains richer basis atoms allowing more meaningful and stable representations of source images. By doing so, SR-based fusion methods generally outperform the traditional MST image fusion methods in both subjective and objective tests.
2. The traditional MST fusion methods are implemented in a multiscale manner, where the selection of the MST decomposition level becomes thereby crucial and tricky. To ensure spatial details can be extracted from the source images, the decomposition level is often set too large. In this case, one coefficient in the low-pass band has a great impact on a large set of pixels in the fused image. Accordingly, an error in the low-pass subband (mainly caused by noise or mis-registration between the source images) will lead to serious artificial effects [19]. The fusion of the high-pass sub-band coefficients is also sensitive to noise and mis-registration in this case. Consequently, the MST fusion methods are generally sensitive to mis-registration, impending their usage in the practical applications where a perfect spatial alignment of different source images is unachievable. In contrast, the SR-based fusion methods are generally implemented in a patch way. More specifically, the source images are first divided into a number of patches of the same size, and the fusion is carried out at the patch level. Moreover, in order to reduce block artifacts and improve the robustness against mis-registration, a sliding window with a step length equal to a fixed number of pixels (e.g., one pixel) is often used in the SR-based fusion methods. In other words, these patches overlap by a fixed number of pixels along the horizontal and vertical directions. Generally, SR-based fusion methods are more robust to mis-registration than MST-based ones.

1.1. SR Image fusion in a nutshel

Since Yang and Li [18] took the first step in applying the SR theory to the image fusion field, a number of SR-based image fusion methods have been proposed. As shown in Fig. 1, the growing appeal of this research area can be observed from the steady increase in the number of scientific papers published in academic journals and magazines since 2010.

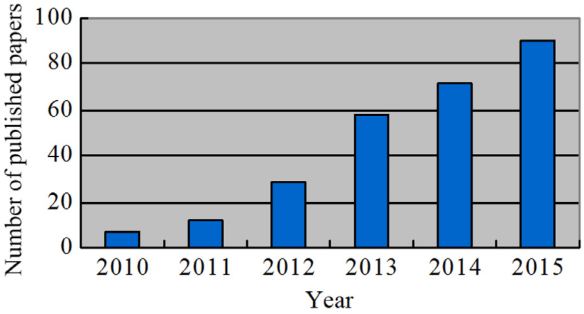

Fig. 1. Numbers of publications on SR-based fusion methods, obtained from the Web of Science indexing service.

The basic idea behind SR-based image fusion is that image signals can be represented as a linear combination of a “few” atoms from a pre-learned dictionary, and the sparse coefficients describe the salient features of the source images. As shown in Fig. 2, the main steps in most SR-based image fusion methods include: (a) segment the source images into some overlapping patches and rewrite each of these patches as a vector; (b) perform sparse representation on the source image patches using pre-defined or learned dictionaries; (c) combine the sparse representations by some fusion rules; (d) reconstruct the fused images from their sparse representations.

The dictionaries employed in these methods may be directly obtained from some fixed (e.g., DCT and Wavelet) basis [18]. They can also be learned from a set of auxiliary images (*global trained dictionary*) [22] or from the input images themselves (*adaptively trained dictionary*) [23] using some learning methods, such as K-SVD [24]. Sometimes, a pair of coupled dictionaries are even simultaneously learned from a high-spatial-resolution image and its spatially-degraded version. Using the coupled dictionaries allows to produce a fused image with higher spatial-resolution [25,26].

Different sparse representation models have been used in image fusion methods. They include: (1) the traditional SR model [15] in which the sparsity constraint (using l_0-norm or l_1-norm) is performed on the representation coefficients; (2) the non-negative SR model [27] in which the sparsity and non-negativity constraints are jointly imposed on the representation coefficients; (3) the robust SR model [28] in which the sparsity constraint is imposed on the reconstruction errors as well as on the representation coefficients; (4) the group-sparsity SR model [29] in which the nonzero representation coefficients are forced to occur in clusters (called group-sparsity) rather than appear randomly; (5) the joint-sparse representation (JSR) model [30] which indicates that different signals from various sensors of the same scene form an ensemble. All signals in one ensemble have a common sparse component, and each employs an individual sparse component.

When fusing the source image patches, the l_1- or l_2-norm of the representation coefficients [18] is generally used. It could possibly benefit from other information to calculate the activity level [9], which measures the information contained in these representation coefficients that is deemed useful during the fusion. Statistical characteristics, such as the sparseness level [27] of their representation coefficients, might also be employed to determine the activity level during the fusion. The energy of the sparse reconstruction errors [28] has been used to determine the activity level when fusing multi-focus images. With an activity calculation defined, a maximum-selecting or a weighted-averaging fusion rule can be employed to directly combine source image patches or indirectly combine representation coefficients of the source image patches [9]. If the representation coefficients are to be combined, the fused image is reconstructed using the pre-learned dictionary and the combined representation coefficients (called the transform-domain fusion method) [27,29–31]. Otherwise, the fused image can be directly obtained from the source image patches according to their

[1] As discussed later, parts of the SR-based fusion methods belong to the spatial-domain-based methods.

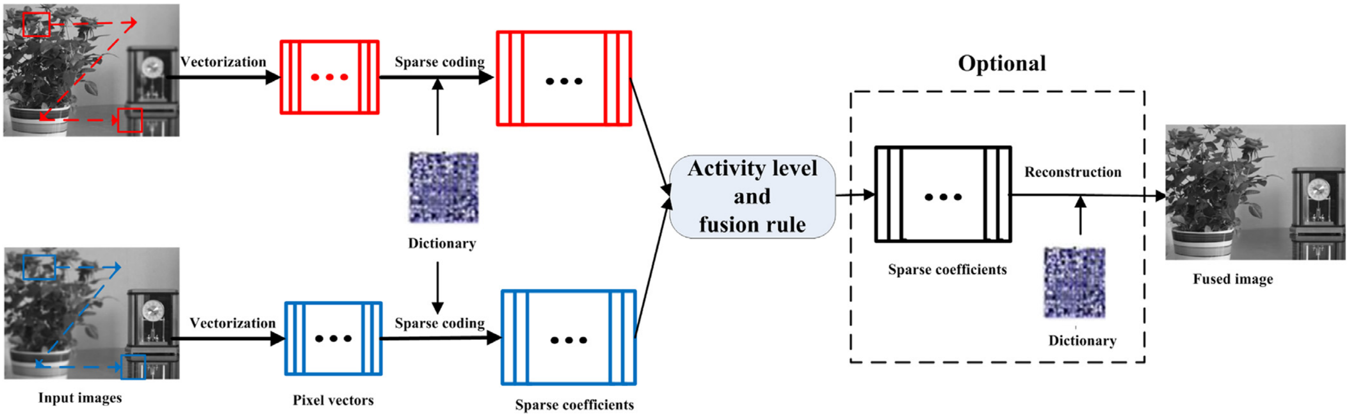

Fig. 2. Diagram of the SR-based image fusion method. (Credit to [2]).

activity level (called the spatial-domain fusion method) [23,28]. The preferred approach depends on the specific intended applications (e.g., fusion of multi-focus images or multi-modality images).

Based on the above analysis, in this paper we will review sparse representation (SR) image fusion methods from the following four key aspects: (1) sparse representation models; (2) dictionary learning methods; (3) activity levels and fusion rules; and finally, (4) applications to multi-focus images and multi-modality (e.g., infrared and visible) image fusion.

1.2. Why this survey?

As pointed out previously, multi-sensor image fusion has always been a hot research topic in the area of image processing, and a considerable number of publications emerge every year. The early reviews [2,5,9,32,33] that focus mainly on traditional MST-based [2,9] or spatial-domain-based (e.g., patches) fusion methods are outdated as they missed out on important recent advances, such as SR-based image fusion methods. In addition, most of them are only limited to one single application of image fusion, such as multi-focus [9], medical [5] or remote sensing image fusion [32,33]. On the other hand, in this paper, we will thoroughly discuss the SR-based fusion methods as well as their applications to fusion of both multi-focus and multi-modality images. Recently, some review papers have also appeared on sparse representation theory [34,35] with the aim to explain the mathematical and theoretical aspects of SR models, but they do not particularly discuss image fusion problems. To the best of our knowledge, there are no previous papers where SR-based fusion methods are reviewed and evaluated. Therefore, it is desirable to put a thorough survey concerning SR-based image fusion in place, which may be useful to a variety of audience, ranging from image fusion learners intended to quickly grasp the current progress in this research area as a whole, to image fusion practitioners interested in applying SR methods to their own problems.

1.3. Paper outlines

The rest of this paper is organized as follows. The available SR models are thoroughly reviewed in Section 2. In Section 3, dictionary learning methods are surveyed. In Section 4, the activity level calculations and fusion rules exploited in the literature with different applications are discussed. In Section 5, the impact of the choice of the components presented in Sections 2–4 on the fusion performance is examined. Finally, conclusions and suggestions for future work are provided in Section 6. Fig. 3 summarizes the structure of this paper.

Table 1
List of vector and matrix related notations.

Symbols	Definition
$x(i)$	The ith entry of the vector x
$X(i, j)$	The (i, j)th entry of the matrix X
$\|x\|_0$	l_0-norm of the vector x, i.e., the number of nonzero entries in the vector x
$\|x\|_1$	l_1-norm of the vector x, $\|x\|_1 = \sum_i \|x(i)\|$
$\|x\|_2$	l_2-norm of the vector x, $\|x\|_2 = \sqrt{\sum_i x^2(i)}$
$\|X\|_0$	l_0-norm of the matrix X, i.e., the number of nonzero entries in the matrix X
$\|X\|_1$	l_1-norm of the matrix X, $\|X\|_1 = \sum_{i,j} \|X(i,j)\|$
$\|X\|_F$	Frobenius -norm of the matrix X, $\|X\|_F = \sqrt{\sum_{i,j} X^2(i,j)}$
$\|X\|_{2,1}$	$l_{2,1}$-norm of the matrix X, $\|X\|_{2,1} = \sum_j \sqrt{\sum_i X^2(i,j)}$
$(\cdot)^T$	Transpose of a vector or a matrix
$X\dagger$	Pseudo inverse of the matrix X

1.4. Notations

We assume that the reader has some basic knowledge of linear algebra and optimization theories. Throughout the paper, a vector is denoted by a low-case letter. A matrix is denoted by a capital letter. All the elements in a vector or a matrix are real-valued. Given a vector x and a matrix X, some notations related to them used in this paper are listed in Table 1.

2. Sparse representation models

Since the traditional SR model [15] was first applied to multi-sensor image fusion, many of its extensions have also been applied to image fusion. For example, a non-negative sparse representation (NNSR) model was introduced for image fusion in [27]. Unlike the traditional SR model that just imposes the sparsity constraint on the representation coefficients, the NNSR model imposes the joint sparsity and non-negativity constraints on the representation coefficients. From the image patch encoding point of view, the interpretation of NNSR model is more intuitive than the traditional SR model.

Assuming the imaging sensors observe the same scene, the source images captured by these sensors are expected to possess common (or redundant) and complementary (distinct) features. Such ideas map well into the joint sparse representation (JSR) model [30], in which all the each sensor image from the same ensemble is automatically decomposed into a common component that can be shared by all the images and an innovation component that describes individual differences. As a result, the JSR model attracts more attention in image fusion, especially in multi-modality image fusion.

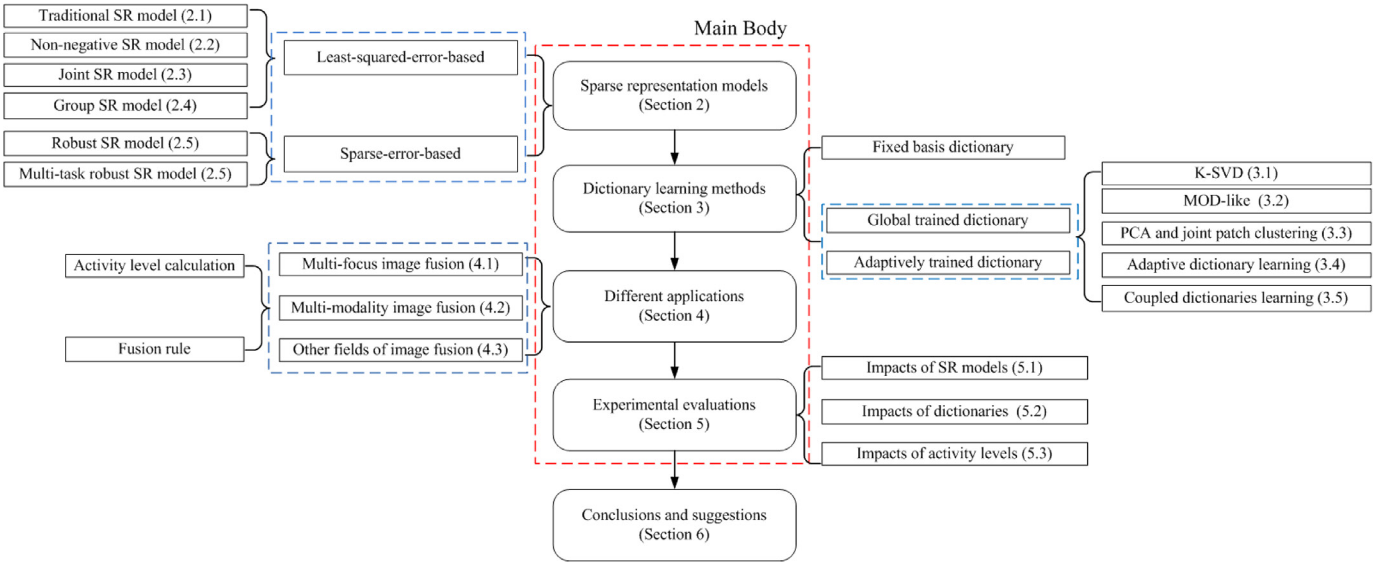
Main Body
Traditional SR model (2.1)
Non-negative SR model (2.2)
Joint SR model (2.3)
Group SR model (2.4)
Robust SR model (2.5)
Multi-task robust SR model (2.5)
Least-squared-error-based
Sparse-error-based
Sparse representation models (Section 2)
Dictionary learning methods (Section 3)
Different applications (Section 4)
Experimental evaluations (Section 5)
Conclusions and suggestions (Section 6)
Fixed basis dictionary
Global trained dictionary
Adaptively trained dictionary
K-SVD (3.1)
MOD-like (3.2)
PCA and joint patch clustering (3.3)
Adaptive dictionary learning (3.4)
Coupled dictionaries learning (3.5)
Activity level calculation
Fusion rule
Multi-focus image fusion (4.1)
Multi-modality image fusion (4.2)
Other fields of image fusion (4.3)
Impacts of SR models (5.1)
Impacts of dictionaries (5.2)
Impacts of activity levels (5.3)

Fig. 3. Organization of this paper.

In [28], a robust sparse representation (RSR) model was introduced to extract the detailed information in a set of multi-focus input images. The RSR model replaces the conventional least-squared reconstruction error with a so-called sparse reconstruction error. By using RSR, any multi-focus image can be decomposed into a fully-defocus image and a sparse but detailed image denoted by the sparse reconstruction error. Distinct from traditional SR-based fusion methods, the reconstruction errors are employed instead of the usual sparse representation coefficients to guide the fusion process. Superiority over the latter SR-based methods is verified in the experimental results.

In this section, we will review some SR models that have been applied in multi-sensor image fusion. We will start by introducing some specific concepts related to sparse representation, so that the reader can understand the basic concepts associated with this theory. Then we will extend these concepts to some more complex representation models.

2.1. Sparse representation (SR) model

The sparse representation model relies on the assumption that many important signals can be represented or approximately represented as a linear combination of a "few" atoms from a redundant dictionary [19,23]. That is, given such a redundant dictionary $D \in R^{n \times M}$ $(n < M)$ containing M prototype n-dimensional signals that are referred to as atoms formed by the columns of the matrix M, a signal $y \in R^n$ can be expressed as $y = Dx$ or $y \approx Dx$. The vector $x \in R^M$ contains the coefficients that represent the signal y in terms of the dictionary D. As the dictionary is redundant, the vector x is not unique. Thus, the SR model was proposed as a method for determining the solution vector x with the fewest non-zero components [23]. Mathematically, this can be achieved exactly assuming negligible noise or inexactly considering noise by solving the optimization problem

$$\min_x \|x\|_0 \ s.t.\ y = Dx, \tag{1}$$

or

$$\min_x \|x\|_0 \ s.t.\ \|y - Dx\|_2^2 \le \varepsilon. \tag{2}$$

The optimization of the above formulas is NP-hard and thus requires approximate techniques, such as the matching pursuit (MP) [36], orthogonal matching pursuit (OMP) [37] or simultaneous OMP (SOMP) [38] algorithms to obtain solutions with low complexity.

Based on recent developments in SR and compressed sensing, the non-convex l_0-minimization problems in (1) and (2) can be relaxed to obtain the convex l_1-minimization problems [15,39] in

$$\min_x \|x\|_1 \ s.t.\ y = Dx, \tag{3}$$

and

$$\min_x \|x\|_1 \ s.t.\ \|y - Dx\|_2^2 \le \varepsilon \tag{4}$$

Solutions can be obtained by using linear programming methods [15,40].

2.2. Non-negative sparse representation (NNSR) model

Considering that properly scaled black and while images can be interpreted as images with positive entries, Wang et al. [27] introduced a non-negative sparse representation (NNSR) model and applied it to the fusion of infrared and visible light images. Different from the traditional SR model which only emphasizes the sparsity constraint using l_0-norm or l_1-norm, NNSR jointly imposes the sparsity and non-negativity constraints on the representation coefficients. It can also be seen as an extension of the traditional non-negative matrix factorization [41] which adds a sparsity inducing penalty.

Let $Y = [y_1, y_2, \ldots, y_N]$ be an observed non-negative data matrix[2] of size $n \times N$ representing a set of N source image patches, each column of which is a data vector (i.e., an image patch) $y_i \in R^n$. Then, given a dictionary $D \in R^{n \times M}$ with M non-negative prototype atoms, the NNSR model coefficients can be obtained from

$$\min_{x_i} \sum_{i=1}^{N} \left(\frac{1}{2} \|y_i - Dx_i\|_2^2 + \lambda \|x_i\|_1 \right), \quad s.t.\ D \ge 0, x_i \ge 0, i = 1, 2, \ldots, N \tag{5}$$

where $x_i \in R^M$ denotes the representation coefficients of the data y_i. Here, owing to the non-negativity, the l_1-norm of the vector x_i is also calculated as the sum of the components in the vector x_i.

[2] Here, a matrix $D = [d_{i,j}]$ is called non-negative if each of its elements $d_{i,j}$ is non-negative. For simplicity, a non-negative matrix D is denoted by $D \ge 0$.

λ refers to the regularization parameter. When $\lambda = 0$, NNSR is reduced to the non-negative matrix factorization. This problem can be simply and efficiently solved by the non-negative sparse coding algorithm [42].

Similar to the traditional SR model, NNSR can also encode the source images efficiently by using a few "active" components. In contrast, the non-negativity constraint makes the representation purely additive (allowing no subtractions), thus enabling NNSR to achieve an easy or intuitive interpretation of the encodings of the source images [27].

2.3. Joint sparse representation (JSR) model and a generalized version

The term "Joint Sparsity", that is, the common sparsity of the entire signal ensemble, was first introduced in [43]. Three joint sparsity models (JSMs) for different situations were presented, **JSM-1** (sparse common component + innovations), **JSM-2** (common sparse supports) and **JSM-3** (non-sparse common +sparse innovations). When different imaging sensors observe the same scene, the source images captured by the sensors are generally expected to possess both "common (or correlated)" and "innovation (or complementary)" information. Accordingly, it is not surprising that JSM-1 has been shown to be more suitable for many image fusion applications, especially for the fusion of multi-modality images [30], when compared with JSM-2 and JSM-3.

In the JSM-1 (or JSM[3]) model, all signals share a common component while each individual signal contains an innovation component. Let $Y_k \in R^{n \times L}$ $(k = 1, 2, \ldots, K)$ denote the L signals of dimension n from the kth sensor which can be represented using [30]

$$Y_k = Y^C + Y_k^U = DX^C + DX_k^U, k = 1, 2, \ldots, K, \tag{6}$$

where $Y^C = DX^C$ denotes the common component for all signals, and $Y_k^U = DX_k^U$ denotes the innovation component for the kth individual signal. $D \in R^{n \times M}$ $(n < M)$ is an over-complete dictionary. X^C and $X_k^U \in R^{M \times L}$ are the sparse coefficient matrices for the common and innovation components, respectively.

Let

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_K \end{bmatrix} \in R^{nK \times L}, \tag{7}$$

$$\underline{D} = \begin{bmatrix} D & D & \mathbf{0} & \cdots & \mathbf{0} \\ D & \mathbf{0} & D & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D & \mathbf{0} & \mathbf{0} & \cdots & D \end{bmatrix} \in R^{nK \times (K+1)M}, \tag{8}$$

$$X = \begin{bmatrix} X^C \\ X_1^U \\ \vdots \\ X_K^U \end{bmatrix} \in R^{(K+1)M \times L}, \tag{9}$$

where $\mathbf{0} \in R^{n \times M}$ is a matrix of zeros. Under the assumed sparseness, the coefficients of JSM model can be computed using [30,44,45]

$$\min_X \|X\|_0 \ s.t. \ \|Y - \underline{D}X\|_F^2 \le \varepsilon, \tag{10}$$

where $\varepsilon \ge 0$ is the error tolerance. Similar to solving (3) in the traditional SR model, the joint sparse coefficient matrix X of the JSM model in (10) can be obtained by using the previously discussed sparse approximation algorithms (e.g., the OMP algorithm [37]). Fig. 4 illustrates the common and complementary information obtained by using the JSR model,[4] where Fig. 4 (c) contains the common background information acquired by the two sensors, while Fig. 4 (d) and (e) contain the complementary information between the two source images. Especially, the man behind the tree captured by the infrared imaging sensor is clearly displayed in Fig. 4 (e).

Considering that the subspace spanned by the innovation component might not be the same as the subspace spanned by the common component, Zhang et al. [30] presented *a generalized version of the JSM model*. In the generalized JSM model, the signals from one ensemble are assumed to depend on two dictionaries, i.e. the common dictionary $D^C \in R^{n \times M}$ and the innovation dictionary $D^U \in R^{n \times M}$, instead of a single dictionary as in the JSM model. Accordingly, (6) and the dictionary matrix $\underline{D}$ in (8) are extended in the generalized JSM model [30], respectively to

$$Y_k = Y^C + Y_k^U = D^C X^C + D^U X_k^U, k = 1, 2, \ldots, K, \tag{11}$$

$$\underline{D} = \begin{bmatrix} D^C & D^U & \mathbf{0} & \cdots & \mathbf{0} \\ D^C & \mathbf{0} & D^U & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D^C & \mathbf{0} & \mathbf{0} & \cdots & D^U \end{bmatrix} \in R^{nK \times (K+1)M}. \tag{12}$$

According to (10), the generalized JSM model can be solved by using the same methods as those for the traditional SR and JSM models. In [30], the generalized JSM model is shown to be sometimes superior to the JSM model in terms of the ability to extract detailed information from the resulting image representations but with little extra computational complexity.

2.4. Group sparse representation model

Most of the existing SR models mentioned previously assume that the non-zero coefficients appear randomly, and do not consider the intrinsic structure of the signals. For that, Li et al. introduced a group sparse representation (GSR) model [29], in which the cluster structure sparsity prior is incorporated and the non-zero elements are forced to occur in clusters (called group-sparsity), rather than appear randomly.

Let $G = \{G_1, G_2, \ldots, G_g\}$ be a partition of the index set $\{1, 2, \ldots, M\}$, where g is the number of groups. Given a dictionary $D = \left[D_{G_1}, D_{G_2}, \ldots, D_{G_g}\right] \in R^{n \times M}$ where D_{G_i} denotes the sub-dictionary with columns identical to D in group G_i, any signal $y \in R^n$ can be represented as [29]

$$y = Dx = \left[D_{G_1}, D_{G_2}, \ldots, D_{G_g}\right]\left[x_{G_1}^T, x_{G_2}^T, \ldots, x_{G_g}^T\right]^T, \tag{13}$$

where $x = [x_{G_1}^T, x_{G_2}^T, \ldots, x_{G_g}^T]^T \in R^M$ denotes the representation coefficients, and x_{G_i} $(i = 1, 2, \ldots, g)$ are the representation coefficients with respect of the sub-dictionary D_{G_2}. In the GSR model, the sparse representation coefficients are found from

$$\min_x \|x\|_{2,0} \ s.t. \ y = Dx \text{ or } \|y - Dx\|_2^2 \le \varepsilon, \tag{14}$$

where $\|x\|_{2,0} = \sum_{i=1}^{g} I\left(\left\|x_{G_i}\right\|_2\right)$, and $I(\,\cdot\,)$ is an indicator function, i.e.,

$$I\left(\left\|x_{G_i}\right\|_2\right) = \begin{cases} 1, & \text{if} \left\|x_{G_i}\right\|_2 > 0 \\ 0, & \text{otherwise} \end{cases}. \tag{15}$$

Similarly, the non-convex $l_{2,0}$-minimization optimization problem in (14) can be relaxed by solving the following convex $l_{2,1}$-minimization problem in (16)

$$\min_x \|x\|_{2,1} \ s.t. \ y = Dx \text{ or } \|y - Dx\|_2^2 \le \varepsilon, \tag{16}$$

[3] In the remaining discussion, the symbol "JSM" denotes the JSM-1 model for simplicity unless expressly specified otherwise.

[4] The test images in Figs.4, 10 and 12 are downloaded from www.imagefusion.org.

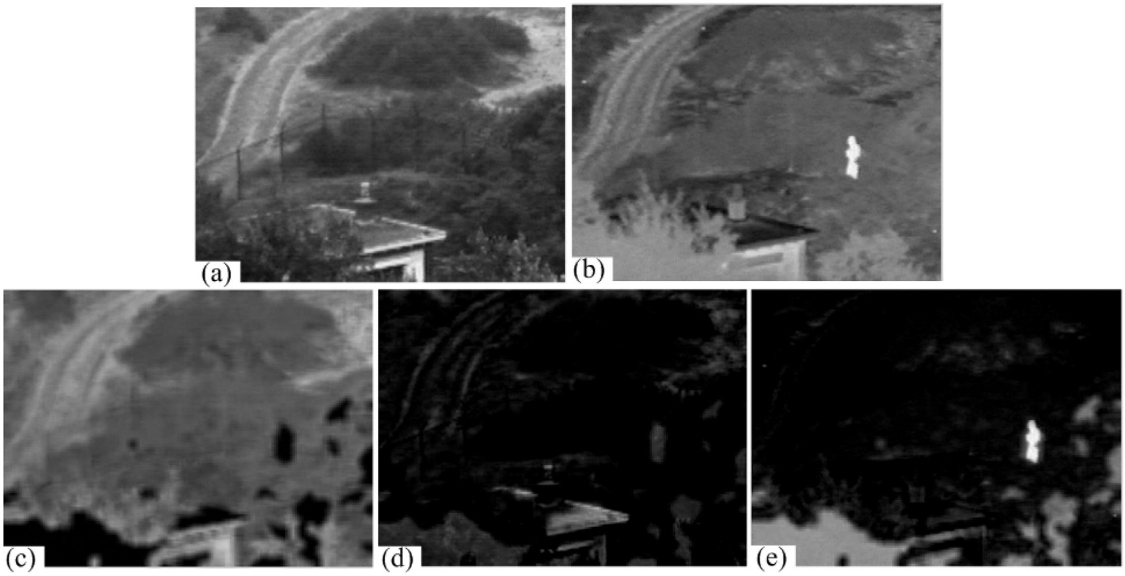

Fig. 4. Illustration of the common and innovation information obtained by using the JSR model. (a) and (b) test images captured by two different sensors; (c) The common component between the two test images; (d) and (e) The innovation components of the test images in (a) and (b), respectively.

where $\|x\|_{2,1}=\sum_{i=1}^{g}\left\|x_{G_i}\right\|_2$. The GSR model can be effectively solved via the Group Orthogonal Matching Pursuit (GOMP) algorithm [46].

Fig. 5 illustrates the representation coefficients obtained by using the SR model and the GSR model. In the GSR model, a dictionary containing 8 sub-dictionaries (i.e., $g = 8$ in (13)) is employed. As shown in Fig. 5(b) and (d), the coefficients obtained by using SR model are sparsely and randomly distributed along the entire horizontal axis. In contrast, the coefficients obtained by using the GSR model are just sparsely located at a few segments along the horizontal axis as shown in Fig. 5(c) and (e). This demonstrates that each local patch can be well reconstructed by using only a few sub-dictionaries, instead of a few random dictionary atoms, in the GSR model.

2.5. Robust sparse representation (RSR) model and a multi-task version

As discussed previously, the traditional SR, NNSR, JSR and GSR models are seen to impose either an l_0-norm or l_1-norm minimization on the representation coefficients to achieve a sparse representation of a signal, while imposing an l_2-norm minimization on the reconstruction errors (e.g., the component $\frac{1}{2}\|y_i - Dx_i\|_2^2$ in (5)).[5] These approaches work well for signals with small levels of Gaussian noise. However, if the signal contains non-Gaussian noise or is corrupted by sparse but strong “outliers”, it may not be possible to achieve a satisfactory result [15].

In [28], Zhang and Levine presented a robust sparse representation (RSR) model by imposing sparse constraints on the reconstruction errors as well as on the representation coefficients. More specifically, let $Y = [y_1, y_2, \ldots, y_N]$ be an observed data matrix of size $n \times N$, each column of which is a data vector $y_i \in R^n$. Further, suppose the observed data Y is partially corrupted by errors or noise $E \in R^{n \times N}$. Then, given a dictionary $D \in R^{n \times M}$ with M prototype atoms, the coefficients of the RSR model are assumed to follow [28]

$$\min_{X,E} \|X\|_1 + \lambda\|E\|_{2,1} \ s.t.\ Y = DX + E, \tag{17}$$

where the matrix $X \in R^{M \times N}$ denotes the sought after matrix of coefficients, and each of its columns $x_i \in R^M$ denotes the sparse coefficient vector for the data y_i. $\lambda > 0$ is a parameter and is used to balance the effects of the two components in (17). The optimization problem in (17) is convex and can be solved by various methods. In [28], Zhang and Levine used the linearized alternating direction method with adaptive penalty (LADMAP) [47,48] to solve this problem because of its high efficiency.

Here, we perform an experiment to demonstrate the robustness of the RSR model to non-Gaussian noise or sparse “outliers”. Similar to [15], we select half of the images in the Extended Yale B database for training and the rest for testing. In the experiment, the pixel intensities of the original images are used as features and stacked as columns of the dictionary matrix D and the data matrix Y. Then the representation coefficient matrix X and reconstruction matrix E are obtained by solving (17).

As shown in Fig. 6, the images reconstructed by the RSR model are superior to those reconstructed by the traditional SR model. For example, there are some ghosts near the eye regions labeled by a green rectangle in Fig. 6(b1) reconstructed using the traditional SR model. This phenomenon looks more severe in Fig. 6(b2). In contrast, these ghosts are greatly reduced in the images reconstructed by the RSR model, as shown in Fig. 6(c1) and (c2). This also demonstrates that the RSR model is more robust to non-Gaussian noise or sparse “outliers” than the traditional SR model.

In order to effectively extract and utilize multiple features for each local image patch during the fusion process, Zhang and Levine generalized the RSR model to multi-task sparsity pursuit and presented a multi-task RSR (MRSR) model [28]. In MRSR, the multi-task sparsity pursuit is achieved by enforcing a joint sparsity constraint on the reconstruction errors across all the tasks.

Let $Y_k = \left[y_{k,1}, y_{k,2}, \ldots, y_{k,N}\right] \in R^{n_k \times N}$ $(k = 1, 2, \ldots, K)$ consist of K feature matrices for K different types of features. The vector $y_{k,i} \in R^{n_k}$ denotes the kth type of feature of dimension n_k for the ith image patch. Correspondingly, the columns $y_{k,i} \in R^{n_k}$ $(k = 1, 2, \ldots, K)$ in these matrices with the same index i and different k denote different types of features for the same ith image patch. N denotes the total number of patches in the image to be considered. Then

[5] In fact, the problems in (3) and (4) are equivalent to the following problem: $\min_x \frac{1}{2}\|y - Dx\|_2^2 + \lambda\|x\|_1$. Thus, the traditional SR model also imposes an l_2-norm minimization on the reconstruction errors.

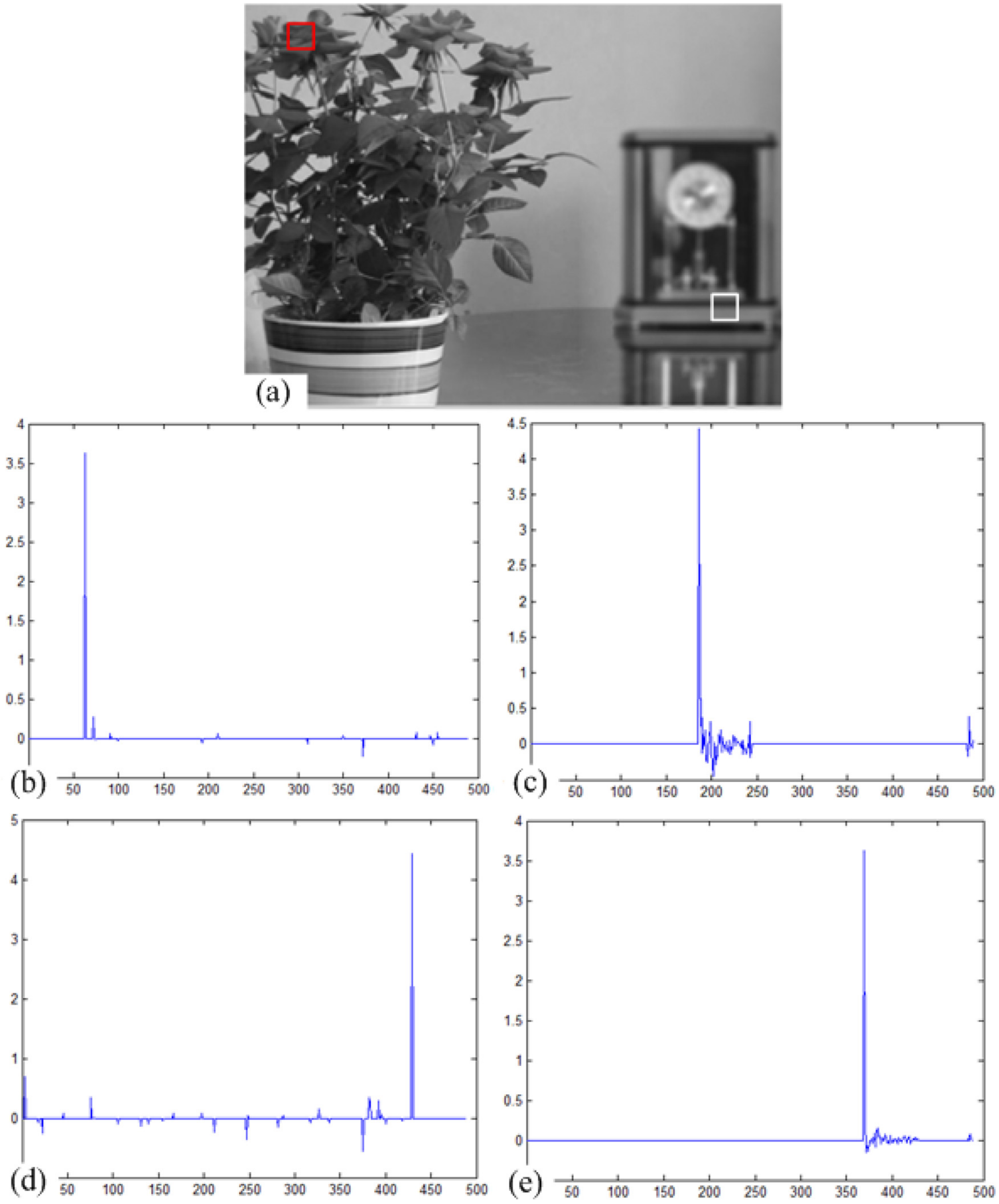

Fig. 5. Illustration of GSR coefficients. (a) Test image; (b) and (c) SR coefficients and GSR coefficients for the red rectangle patch in (a), respectively; (d) and (e) SR coefficients and GSR coefficients for the white rectangle patch in (a), respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the MRSR coefficients are assumed to satisfy [28]:

$$\begin{aligned} &\min_{X_k,E_k} \sum_{k=1}^{K} \|X_k\|_1 + \lambda \|E\|_{2,1} \quad , \\ &s.t.\ Y_k = D_k X_k + E_k,\ \ k = 1, 2, \ldots, K \end{aligned} \tag{18}$$

where $D_k \in R^{n_k \times M_k}$ is a dictionary with M_k prototype atoms for the kth type of feature. $X_k \in R^{M_k \times N}$ and $E_k \in R^{n_k \times N}$ denote the SR coefficients and the reconstruction errors for the kth feature matrix Y_k, respectively. The joint error matrix E is formed by concatenating the vertical columns of matrices E_1, E_2, ... ,E_K.

As discussed in [28,49], the corresponding columns in the matrices E_1, E_2, ... ,E_K with the same index will be compelled to have similar magnitudes by imposing the $l_{2,\,1}$-norm minimization on the matrix E. As for the RSR model, the optimization problem of MRSR can also be solved using LADMAP [47,48].

2.6. Summary

A close look at the aforementioned algorithms reveals that the essential difference among the SR models discussed above is where they apply the constraints, either on the representation coefficients, the reconstruction errors or on both. It can also be noticed

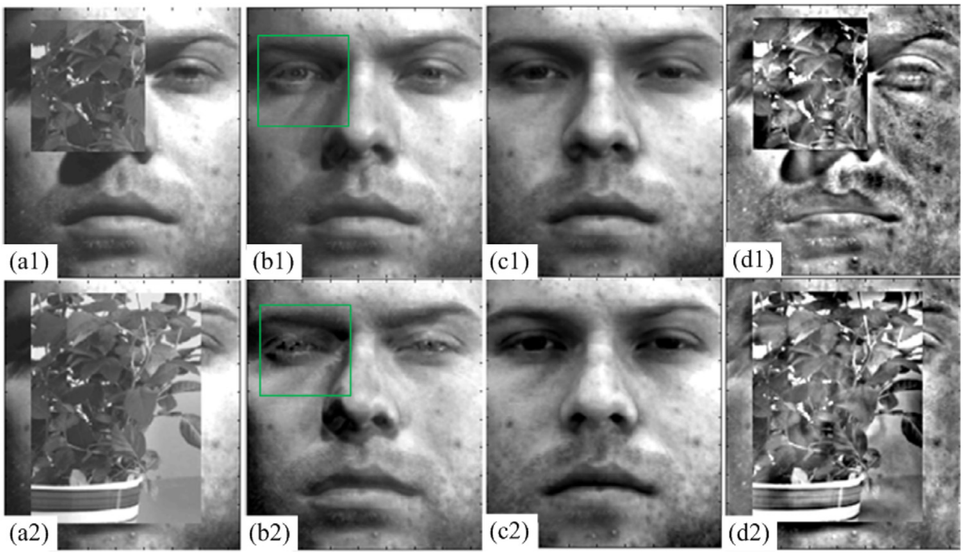

Fig. 6. Reconstructed results for images with occlusions. (a1) and (a2) are occluded test images of the first subject in the Extended Yale B database with 23% and 61% occlusion, respectively; (b1) and (b2) are reconstructed images using the dictionary atoms from the first subject and their corresponding SR coefficients for (a1) and (a2), respectively; (c1) and (c2) are reconstructed images using the dictionary atoms from the first subject and their corresponding RSR coefficients for (a1) and (a2), respectively; (d1) and (d2) indicate the RSR reconstruction errors for (a1) and (a2), respectively.

Table 2
Summary of the sparse representation models employed in multi-sensor image fusion.

Models		Representation coefficients constrains	Reconstruction error constrains
Least-squared-error-based	SR	Sparsity constraint	Least squared minimization constraint
	NNSR	Sparsity and non-negativity constraint	
	JSR	Sparsity common component and innovation components constraint	
	GSR	Group-sparsity constraint	
Sparse-error-based	RSR	Sparsity constraint	Sparsity constraint
	MRSR	Sparsity constraint	Joint sparsity constraint cross error matrices of multiple tasks

that the traditional SR, NNSR, JSR and GSR models impose different constraints on the representation coefficients but the same least squared minimization constraint on the reconstruction errors. These SR models can thus be called *least-squared-error-based* models. Differently, the RSR model and MRSR models replace the conventional least-squared reconstruction error with a so-called sparse reconstruction error. Therefore, the RSR and MRSR models can be called *sparse-error-based* models.

In contrast to those least-squared-error-based SR methods, using the sparse-error significantly improves the robustness of the RSR model against the non-Gaussian noise or sparse but strong corruptions, thereby facilitating practical applications. More importantly, many important features, including the detailed information contained in an image, can be denoted by the sparse error components obtained using the RSR model. Table 2 summarizes the previously mentioned sparse representation models.

Basically, the NNSR, JSR, GSR, RSR, and MRSR models somewhat improve the traditional SR model in various aspects, and they generally perform better than the SR model when applied to multi-sensor fusion applications. However, it is difficult to explain the suitability of a model for a specific application from the general point of view. Instead, we draw the conclusion according to the experimental results, which reveal that the RSR model seems to be more suitable for multi-focus image fusion; the NNSR and JSR are more suitable for multi-modality image fusion; and the GSR model can facilitate both as it achieves generally good results for these two applications. It is necessary to point out that the performance may be further improved if the dictionary of a model complies with the characteristics of the data. That is to say, it does not make sense to expect a universal dictionary that can enhance the performance of all the models. As a result, designing an appropriate dictionary for each model deserves further investigation.

3. Dictionary learning methods in multi-sensor image fusion

Constructing a good dictionary is of fundamental importance for the performance of an SR-based image fusion method. Generally, there are two categories of methods to construct an over-complete dictionary. The first one uses some fixed basis [18,50]. In [18] for instance, an over-complete separable version of the DCT dictionary is constructed by sampling cosine waves with different frequencies. In [50], a hybrid dictionary consisting of a DCT basis, a wavelet 'db1' basis, a Gabor basis and a ridgelet basis is constructed. Employing a fixed basis has the advantages of simplicity and fast implementation. Since this approach is not customized by using appropriate input image data, it may provide inferior performance for certain types of data and applications.

The second category of methods is to construct an over-complete dictionary by using some learning methods, such as PCA, MOD and K-SVD [24]. These methods can be further divided into *global-trained-dictionary-based* [19,22,44,50] and *adaptively-trained-dictionary-based* [23,27,28,30,45], according to their employed training images. In the former methods, a public training database that generally contains many high-resolution images is employed to construct the training data for dictionary learning. For example, in [19], the training data consists of 100,000 8 × 8 patches, randomly sampled from a database of 40 high-quality images. While in the latter methods, the input images are directly used to construct the training data. For example, in [27], the training data for dictionary learning contains 20,000 8 × 8 patches, which are randomly sampled from the source infrared and visible images. In [23], local patches from the input multi-focus images are used as the training samples to learn a dictionary. In [28], the input image patches are directly employed to construct an over-complete dictionary. These dictionaries are adaptive to the input image data and thus have the potential to outperform the commonly used fixed dictionaries. Accordingly, these learned dictionaries are more widely adopted in SR-based image fusion. In the rest of this section, we review some dictionary learning methods used in multi-sensor image fusion.[6]

3.1. Dictionary learning using K-SVD

Let $Y = [y_1, y_2, \dots, y_N] \in R^{n\times N}$ be a training data matrix, where $y_i \in R^n$ is the ith sampled data vector. Our goal is to learn a dictionary $D = [d_1, d_2, \dots, d_M] \in R^{n\times M}$ and a sparse coefficient matrix $X = [x_1, x_2, \dots, x_N] \in R^{M\times N}$, such that the product of D and X can approximate the original data matrix Y efficiently. If X were known, the over-complete dictionary D could be obtained from the matrix Y via solving

$$\min_{D,X} \|Y - DX\|_F^2 \ s.t.\ \|x_i\|_0 \le \tau, i = 1, 2, \dots, N, \tag{19}$$

where τ denotes the upper bound for the number of the non-zero entries in x_i. The solution to (19) for both D and X can be obtained by using the popular dictionary learning algorithm K-SVD [24], which iteratively alternates between two steps: sparse coding (find X) and dictionary updating (find D).

In the sparse coding step, D is assumed to be fixed, and the optimization problem of (19) is reduced to a search for sparse representations with coefficients summarized in the matrix X. For that, the criterion is rewritten as

$$\|Y - DX\|_F^2 = \sum_{i=1}^{N} \|y_i - Dx_i\|_2^2. \tag{20}$$

Therefore, the problem in (19) can be decoupled into N optimization problems of the form

$$\min_{x_i} \|y_i - Dx_i\|_2^2 \ s.t.\ \|x_i\|_0 \le \tau, i = 1, 2, \dots, N. \tag{21}$$

This problem can be efficiently solved by the MP [36] and OMP [37] algorithms mentioned in Section 2.

In the dictionary updating stage, the coefficient matrix X and the dictionary D are both assumed to be fixed. Only one column d_k in the dictionary and the coefficients that correspond to it (i.e., the kth row of X, denoted as x_k^T) are considered each time. For that, the multiplication DX in (19) is decomposed into the sum of K rank-1 matrices. During the updating, K-1 terms are supposed to be fixed and one, i.e., the kth, remains in question. More specifically, the metric in (19) is rewritten as [24]

$$\begin{aligned} \|Y - DX\|_F^2 &= \left\| Y - \sum_{j=1}^{M} d_j x_j^T \right\|_F^2 \\ &= \left\| \left(Y - \sum_{j\ne k} d_j x_j^T \right) - d_k x_k^T \right\|_F^2 = \left\| E_k - d_k x_k^T \right\|_F^2 \end{aligned}, \tag{22}$$

where E_k stands for the error for all the N samples when the kth atom is removed. Minimizing the function in (22) is equivalent to finding a rank-1 matrix that closely approximates the error term E_k in Frobenius norm. The rank-1 matrix is described by the atom d_k and the row vector x_k^T. These can be obtained simply by using singular value decomposition (SVD) on E_k. Moreover, to ensure the sparsity of the vector x_k^T, some modifications are further performed on (22). More details can be found in [24].

3.2. Dictionary learning using MOD-like

In [30], the authors present a dictionary learning method (termed as *MODJSR*) for the JSR model. Similar to the traditional dictionary learning methods using K-SVD, MODJSR is also implemented by alternating the sparse coding stage and the dictionary updating stage. In the second stage, dictionary updating is performed as a problem by the "Landweber" update [51] with an initial point obtained by the method of optimal directions (MOD). This method is shown to have higher computational efficiency than the K-SVD method.

Suppose $Y_k \in R^{n\times L} (k = 1, \dots, K)$ are signals from the same ensemble, i.e., from different source images of the same scene. Motivated by dictionary learning for the standard SR, the dictionary learning method, MODJSR, for the JSR model is defined as [30]

$$\min_{D,X} \frac{1}{2} \|Y - \underline{D}X\|_F^2 \ s.t.\ \|x_t\|_0 \le \tau, t = 1, 2, \dots, L. \tag{23}$$

Here, the data set matrix $Y \in R^{nK \times L}$, the dictionary matrix $\underline{D} \in R^{nK\times(K+1)M}$ and the coefficient matrix $X \in R^{(K+1)M\times L}$ are constructed as in (7), (8), and (9), respectively. τ denotes the maximal number of non-zeros coefficients used in each column of X.

Adopting the block-coordinate descent idea, an alternating strategy is used to solve (23) with two stages. The first stage employs a joint sparse coding. That is, fixing the dictionary D, the joint sparse coefficient matrix X can be obtained by solving (10) via OMP [37] to take advantage of its simplicity and fast execution.

The second stage updates the dictionary. Fixing the joint sparse coefficient matrix X, the dictionary $\underline{D}$ in (23) could be updated simply by $\hat{\underline{D}} = YX^T(XX^T)^{-1}$ with MOD. However, XX^T may not always be full rank. The majorization method could be also directly employed, but it is slow due to using the "Landweber" update which is a gradient update. If the dictionary is updated by the "Landweber" update, the initial point can be obtained by MOD. Then D is found by solving [30]

$$\begin{aligned} \min_D f(D) &= \min_D \frac{1}{2} \|Y - \underline{D}X\|_F^2 \\ &= \min_D \sum_{k=1}^{K} \frac{1}{2} \left\| Y_k - D\left(X^C + X_k^U\right) \right\|_F^2 \end{aligned}. \tag{24}$$

The optimum of the objective function satisfies

$$0 = \frac{d}{dD} f(D). \tag{25}$$

Hence,

$$W = DH, \tag{26}$$

where $W = \sum_{k=1}^{K} Y_k \left(X^C + X_k^U\right)^T$ and $H = \sum_{k=1}^{K} \left(X^C + X_k^U\right) \left(X^C + X_k^U\right)^T$. Since X is sparse, the non-zero elements of H are often concentrated on the diagonal and $H_{ii} \ge 0$ $(i = 1, \dots, M)$, $rank(H) = M$ holds with high probability [52] due to Diagonal Dominance theory. When $rank(H) = M$, the dictionary D is simply updated by $D = WH^{-1}$. Otherwise, it is updated by the "Landweber" rule as [30,51]

$$D^{[k+1]} = D^{[k]} + \frac{1}{\sigma} \left(W - D^{[k]}H\right)H^T, \tag{27}$$

where σ is a constant satisfying $\sigma > \|H^TH\|_F$. A good initial point, obtained by MOD and given by $D^{[0]} = WH^{\diamond}$ is employed while updating the dictionary updating for higher computation efficiency. Here $H^{\blacklozenge}$ is computed as $H^{\diamond} = U\Sigma^{\dagger}U^T$ and the matrices U and Σ result from the SVD of the matrix H, i.e., $H = U\Sigma U^T$.

[6] It should be noted that the methods to be discussed are adopted for the global-trained dictionaries as well as the adaptively-trained dictionaries.

3.3. Dictionary learning using PCA and joint patch clustering

Since the connection of sparsity and clustering was shown to be desirable in image restoration tasks [31,53], some new dictionary learning frameworks combined with clustering of non-local patches were recently presented [54,55]. Motivated by clustering-based dictionary learning techniques, the authors presented an efficient dictionary learning method based on a joint patch clustering for multi-modal image fusion in [31]. This is also the first attempt towards applying clustering-based dictionary learning to image fusion.

Conventional dictionary learning methods using K-SVD, such as the ones discussed in the previous subsections, generally produce redundant or highly structured dictionaries [31]. The proposed dictionary learning in [31] aims to remove the redundancy while maintaining or improving the quality of the multimodal image fusion. Under an assumption that common image structures are distributed across the source images from different sensor modalities, patches from different source images are clustered together according to local structural similarities. Then sub-dictionaries that best describe the underlying structure of each cluster by using only a few principal components are constructed. Finally, these sub-dictionaries are combined to form a final dictionary.

Since each sub-dictionary consists only of a few principal components of each joint patch cluster, the final dictionary constructed ends up with much smaller size than those learned by K-SVD. Although it is more compact, the constructed dictionary still contains the most informative components from each joint patch cluster. As a result, the computational complexity of the subsequent fusion method is greatly reduced while the fusion performance is maintained.

3.4. Dictionary learning for adaptive sparse representation

In the traditional SR models introduced in Section 2, a highly redundant dictionary is always needed to satisfy signal reconstruction requirements since the structures vary significantly across different image patches. However, this may result in potential visual artifacts as well as high computational cost. To address this problem, Liu and Wang [22] introduced an adaptive sparse representation (ASR) model, in which a set of more compact sub-dictionaries are learned from numerous high-quality image patches. These patches have already been pre-classified into several corresponding categories based on their gradient information.

Let $P = \{p_1, p_2, \ldots, p_N\}$ be a training data matrix, where $p_i \in R^n$ is the i-th sampled data or image patch. The patches in set P are first classified into K categories $\{P_k | k = 1, 2, \ldots, K\}$ according to their dominant gradient directions. Then a total of $K+1$ sub-dictionaries $\{D_o, D_1, \ldots, D_K\}$ are obtained, in which D_0 is learned from all the patches in P having no clear dominant directions, whereas $\{D_k | k = 1, 2, \ldots, K\}$ is learned from the patches in each corresponding subset $\{P_k | k = 1, 2, \ldots, K\}$ that have specific dominant directions described by category k. In this method, the dominant gradient direction of each signal y_i is first computed, after which the sub-dictionary D_{k_i} is adaptively selected as the dictionary. An example of the ASR dictionary learning with $K = 6$ is shown in Fig. 7.

3.5. Coupled dictionaries learning

In [56], sparse representation was applied to single image super-resolution. The main idea of the method is to assume that the up-sampled low-resolution (LR) and high-resolution (HR) image patch pairs share the same sparse coefficients with respect to their own dictionaries. Recently, this idea was applied to multi-sensor image fusion [25,57,58] as well as pan-sharpening [26,59].

In order to construct a pair of coupled dictionaries, two training sets for the LR and HR dictionaries are first constructed from the same set of HR training images[7] as shown in Fig. 8 and explained thereafter. Each high-resolution image I is blurred and down-sampled (with a user-defined factor) to generate a LR image. The latter is then up-sampled back to the original size using Bicubic interpolation and the resulting image is seen as a LR image. A pair of training sets $\left\{y_i^H \in R^n | i = 1, 2, \ldots, N\right\}$, $\left\{y_i^L \in R^n | i = 1, 2, \ldots, N\right\}$ are thus created by extracting patches from the original HR image I and its degraded LR version, respectively, in which y_i^H and y_i^L with the same index i correspond to the same spatial position in the HR and LR images. With the assumption that sparse coefficients of the LR image patch y_i^L over the LR dictionary $D_L \in R^{n \times M}$ are the same as those of the HR image patch y_i^H over the HR dictionary $D_H \in R^{n \times M}$, the coupled dictionaries D_H and D_L can be learned by solving the following optimization problem [25]

$$\{D_H, D_L, X\} = \underset{D_H, D_L, X}{\arg\min} \sum_{i=1}^{N} \left\| y_i^H - D_H x_i \right\|_2^2 + \sum_{i=1}^{N} \left\| y_i^L - D_L x_i \right\|_2^2, \quad (28)$$

$$s.t.\ \forall i \|x_i\|_0 \le \tau$$

where $X = [x_1, x_2, \ldots., x_N] \in R^{M \times N}$ is the matrix containing the sparse coefficients, and τ controls the sparsity level. By introducing auxiliary variables $Y^H = [y_1^H, y_2^H, \ldots, y_K^H] \in R^{n \times N}$, $Y^L = [y_1^L, y_2^L, \ldots, y_N^L] \in R^{n \times N}$, $Y = [(Y^H)^T, (Y^L)^T]^T \in R^{2n \times N}$, and $D = [(D_H)^T, (D_L)^T]^T \in R^{2n \times M}$, problem (28) is equivalently transformed to (19) and can thus be efficiently solved by K-SVD.

3.6. Summary

As discussed in this section, many dictionary learning methods have been presented or applied to multi-sensor image fusion. Among these methods, the K-SVD method, thanks to its simplicity and generalization, is the most broadly adopted by the existing SR-based fusion methods. To some extent, the learning procedure of the ASR dictionary and the coupled dictionary are also K-SVD like based on the same principle. It is worthwhile pointing out that each dictionary learning method has its pros and cons, meaning that there is no universal dictionary that suits all applications.

Using these methods, a globally-trained dictionary or an adaptively-trained dictionary can be generated during the fusion process. These learned dictionaries are adaptive to the input image data and usually perform better than the fixed dictionaries in terms of the extraction and representation of significant features in an image. However, these learned dictionaries generally contain a large number of atoms in order to accurately reconstruct an input image patch. This increases the redundancy among the dictionary atoms and thus degrades the subsequent fusion performance to some extent. Moreover, this also increases the computational complexity of a fusion method. In Table. 3, we compare some existing dictionary learning methods with respect to the number of sub-dictionaries, redundancy, applicable model and consumed computation power. Nevertheless, how to learn a dictionary with a fixed small number of atoms and yet maintain a good representation capability for different SR models and fusion applications is desirable and still a challenging problem in multi-sensor image fusion.

4. Applications of different SR-based fusion methods

So far, SR-based image fusion methods have been used in a wide variety of applications, such as multi-focus image fusion, and multi-modality (e.g., infrared and visible light) image fusion. These

[7] For pan-sharpening, the training sets may be constructed from the HR panchromatic source images.

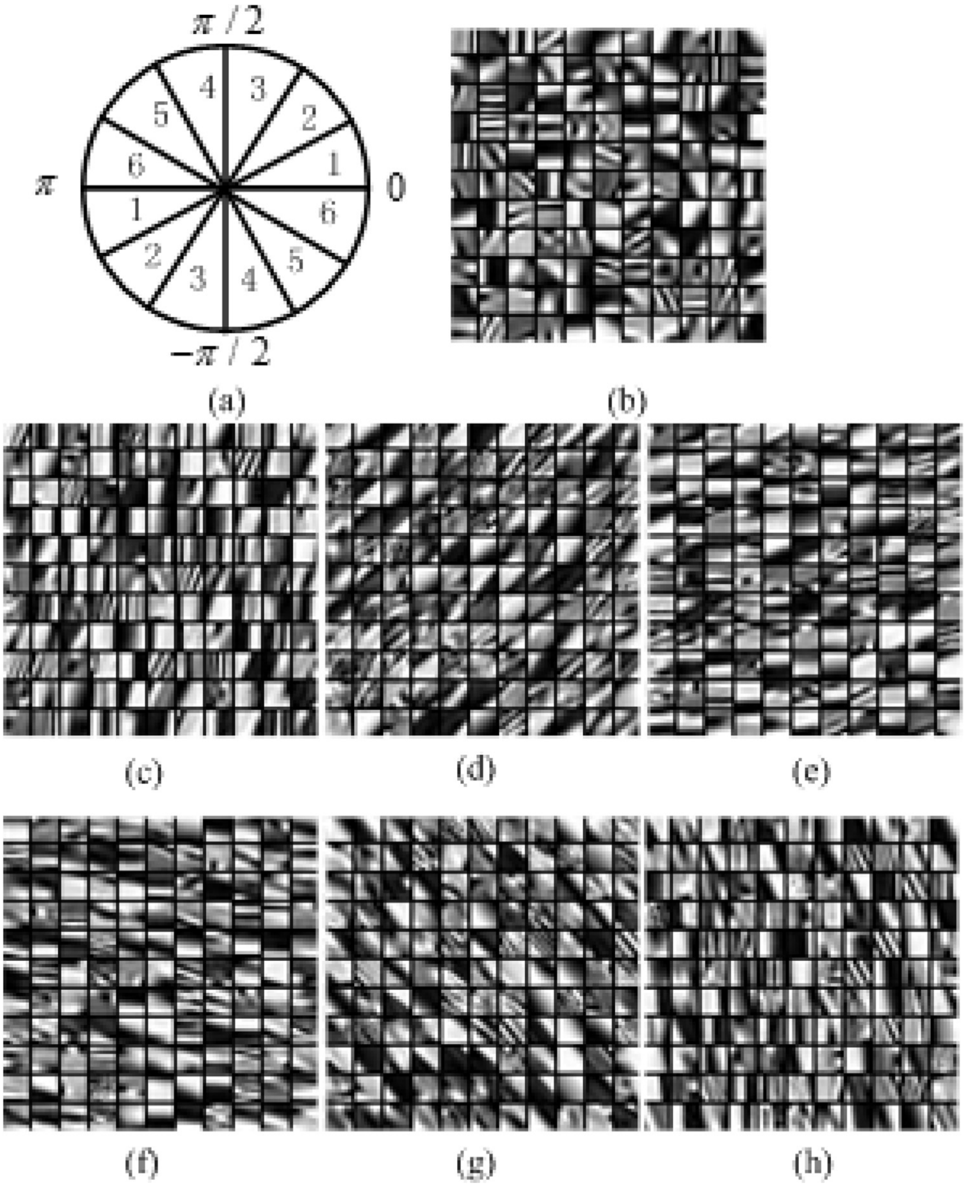

Fig. 7. Learning sub-dictionaries in the ASR model. (a) Illustration of the dominant orientation division; (b)–(h) Learned sub-dictionaries $\{D_k|k = 0, 1, \ldots, 6\}$, respectively. (Credit to [22]).

Table 3
Comparison of different dictionary learning methods.

	Number of dictionaries	Redundancy	Applied model	Computation efficiency
K-SVD-DL	1	High	SR, RSR, MRSR	Low
MOD-DL	1	High	JSR	High
PCA-DL	1 (Multiple sub-dictionaries)	Low	SR, GSR, RSR, MRSR	High
ASR-DL	> 1 (Specific dominant directions) + 1 (common)	Low	SR, RSR, MRSR	Medium
Coupled-DL	2	High	SR, NNSR, RSR, MRSR	Low

applications are targeting different fusion goals, and thus have different fusion strategies. In this section, we will review some applications of SR-based fusion methods for fusing multi-focus images or infrared with visible images.

4.1. Multi-focus image fusion

Due to the limited depth-of-focus of optical lenses in CCD devices, it is often not possible to obtain an image that contains all of the relevant objects in focus. As shown in Fig. 9,[8] this issue can be overcome by multi-focus image fusion, in which several images with different focus points (e.g., Fig. 9(a) and (b)) are combined to form a composite image (e.g., Fig. 9(c)) with full-focus. The basic requirement for multi-focus image fusion is that only the focused regions should be extracted from the given multi-focus input im-

[8] The test images in Fig. 9 and the soon Fig. 11 are downloaded from http://home.ustc.edu.cn/~liuyu1.

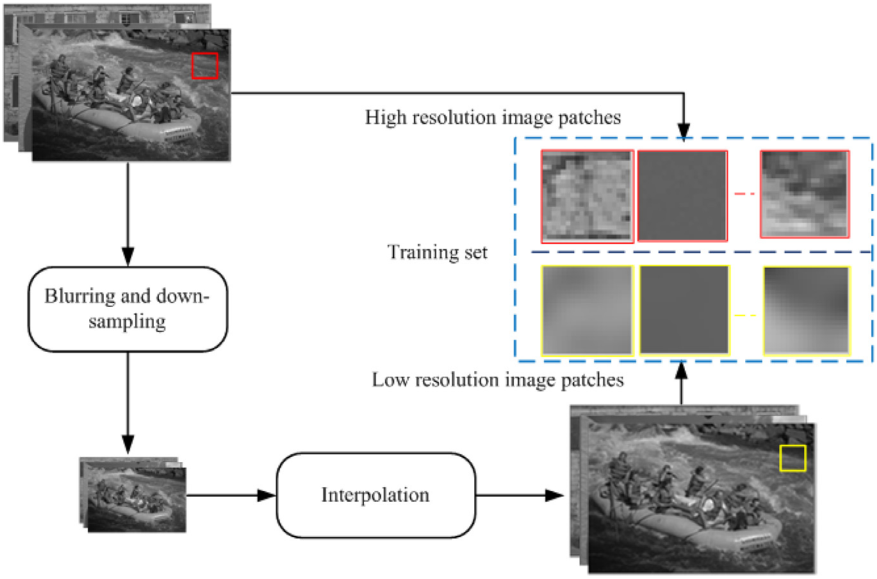
High resolution image patches
Training set
Blurring and down-sampling
Low resolution image patches
Interpolation

Fig. 8. Procedure to construct the training sets for the coupled dictionaries.

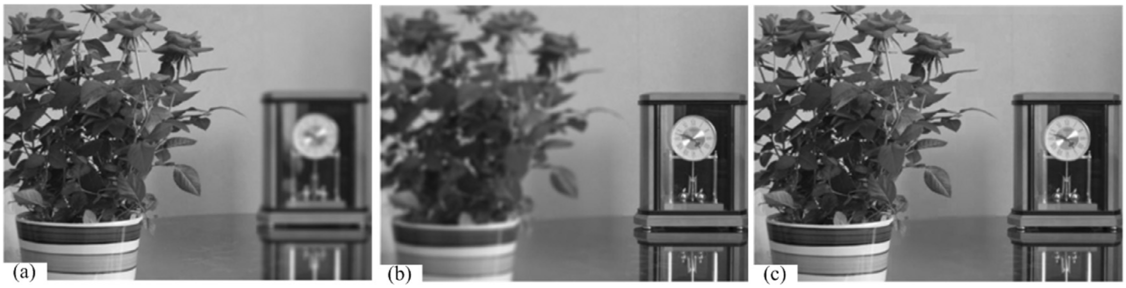

Fig. 9. Illustration of multi-focus image fusion. (a) Focus on the flower; (b) Focus on the clock; (c) Fused image with full-focus.

ages and then preserved in the fused image, while all of the defocused regions should be discarded.

As shown in Fig. 2 in Section 1, the SR-based multi-focus image fusion generally involves the following steps: (1) Divide the source images into a larger number of image patches of the same size (e.g., 8 × 8). In order to reduce block artifacts and improve robustness to mis-registration, a sliding window at a step length of a fixed number of pixels (e.g., one pixel) is also often used in this step. That is to say, these patches overlap by a fixed number of pixels along the horizontal and vertical directions, respectively. (2) Reorder each of these patches as a vector of n-dimensions (e.g., $n = 8 \times 8 = 64$). (3) Sparsely code these vectors via different SR models and pre-constructed dictionaries introduced in Sections 2 and 3. The traditional SR model introduced in Section 2.1 is the most widely used in multi-focus image fusion. The dictionaries directly learned from a set of training images with high-resolution using K-SVD are also the most popular in these methods. (4) Define activity levels and then construct the fused image with different fusion rules.

Activity level reflects the importance of each local image patch. Particularly, for multi-focus image fusion, the activity level should reflect the focus information of each image patch. In SR-based multi-focus image fusion methods, the activity level is generally defined as the l_0-norm, l_1-norm or the l_2-norm of the sparse coefficient vector for each image patch, i.e.,

$$A(p_{k_i}) = \left\| x_{k_i} \right\|_j \tag{29}$$

where p_{k_i} denotes the ith patch from the kth source image, x_{k_i} denotes the representation coefficient vector corresponding to the patch p_{k_i}, and $j = 0$, 1, or 2 describes which norm function is employed to define the activity level.

Sometimes, relatively more sophisticated activity levels are also defined. For example, in [23], the correlation between the sparse representation of the input images and the pooled features obtained in the previous dictionary learning phase is used as the decision map for the fusion. As opposed to most SR-based multi-focus image fusion methods employing the sparse representation coefficients to define activity level, the fusion method presented in [28] employs the sparse reconstruction error, more specifically, the l_2-norm of each column vector in the sparse error matrix obtained by the RSR model, to define the activity level for each source image patch.

There are two different ways to construct the fused image after the activity level of each image patch is determined. Accordingly, different SR-based multi-focus image fusion methods are divided into two categories, *transform-domain-based* and *spatial-domain-based*. In the transform-domain-based fusion methods [18,22,29,50,60–64], the representation coefficients of fused image patches are first obtained from the corresponding representation coefficients of source image patches according to their activity levels. Then the fused image patches are constructed by multiplying the pre-defined dictionary with the obtained representation coefficients. On the other hand, in the spatial-domain-based fusion methods [23,28], the fused image patches are directly extracted from the source image patches according to their activity levels.

In general, both the maximum-selection and weighted-averaging fusion rules (or fusion strategies) might be employed to

Table 4
Some state-of-the-art SR-based multi-focus image fusion methods.

Method		Model	Dictionary	Fusion rule
Transform-domain-based	[18,50,60,61]	SR	Learned from a set of images [18,50,61] Fixed DCT basis [18,50,60] Fixed hybrid basis [50] Fixed hybrid basis [50]	Maximum l_1-norm selection of representation coefficient vectors [18] Maximum selection of absolute coefficient vector entries [50] Maximum l_2-norm selection of representation coefficient vectors [60] Weighed averaging of representation coefficient vectors [61]
	[29]	Group SR	Learned from a set of images	Maximum l_2-norm selection of representation coefficient vectors
	[22]	Adaptive SR	Multiple dictionaries with different dominant directions learned from a set of images	Maximum l_1-norm selection of representation coefficient vectors
	[62,63]	JSR	Learned from source images	Summing of representation coefficient vectors
	[64]	Extended JSR	Learned from a set of images	Maximum l_1-norm selection of representation coefficient vectors
	[23]	SR	Learned from source images	Maximum correlation between the sparse representations of input source images and the training pooled features
Spatial-domain-based	[28]	RSR	Data itself	Maximum l_2-norm selection of sparse reconstruction error vectors
	[28]	Multi-task RSR	Data itself	Maximum l_2-norm selection of joint sparse reconstruction error vectors

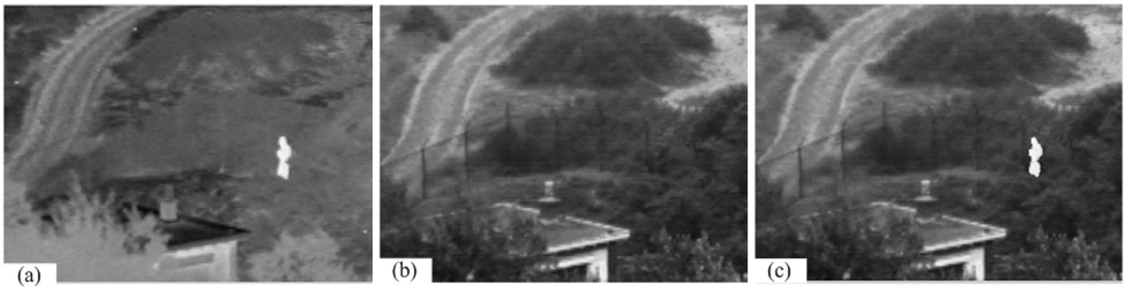

Fig. 10. Illustration of infrared and visible image fusion. (a) Infrared image; (b) Visible light image; (c) Fused image.

determine the fused image patches or their representation coefficients. However, in the SR-based multi-focus fusion methods, the maximum-selection fusion rule is more popular. In this approach, the fused image patch or its sparse representation is generally selected from the input image patch or its sparse representation with the highest activity level. Some state-of-art SR-based multi-focus image fusion methods are summarized in Table 4.

4.2. Multi-modality image fusion

It is becoming more common to employ multiple types of imaging sensors in video surveillance to improve the robustness, in which visible light and infrared imaging sensors are normally combined. Image fusion allows the information captured by these different sensors to be sufficiently and effectively integrated to create a composite image, containing more useful information than any of the individual input images. This image can be used to better interpret the scene [3]. Multi-modality image fusion has also been widely applied to many other fields such as medical imaging.

A video surveillance application is shown in Fig. 10 (a), where the moving person is evident in the image taken by the infrared video camera. However, the scene environment (e.g., the hedges and the shrubs) is better displayed in the visible-light image (Fig. 10(b)), in which the moving targets are difficult to see. By fusing the two input images, the moving target from the infrared camera and the background scene (or the environment) from the visible light camera are well integrated. As shown in Fig. 10(c), the fused image clearly shows that there is a man in the scene.

SR has also been applied to multi-modality image fusion, including infrared and visible light sensors [19,27,30,31,44,45,65]. Due to different imaging technologies of the sensors, these multi-modality images of the same scene captured by different image sensors provide redundant and complementary information. The basic job of a multi-modality image fusion approach is to properly employ the redundant and complementary information available from the different input images [66].

Interestingly, this notion maps well into the JSR model and this is reflected by the fact that, in addition to the traditional SR model, the JSR model is popular in multi-modality image fusion [30,44,45,70]. The reason for this is that in the JSR model, all the signals from the same ensemble are automatically decomposed into a common component that is shared by all the signals and an innovation component that describes each individual signal. The common component describes the redundant information among all the signals, while the innovation component describes the complementary information [45]. Accordingly, JSR already extracts the required information needed for fusion. In the subsequent fusion phase, the innovation components for the input images are combined together by using a weighted-averaging [30,45] or a summing [44,70] fusion strategy. The final fused image is obtained by integrating the common component shared by all the input images into the previously combined innovation component.

Finally, it should be noted that almost all SR-based multi-modality image fusion methods are transform-domain-based. This may result from the fact that patches from the multi-modality input images corresponding to the same spatial positions have greatly diverse characters because of the different sensor technologies. Subsequently, many spatial artifacts will be introduced during the fusion if a spatial-domain-based method is adopted which tends to produce higher activity levels. Alternatively, a transform-domain-based method may reduce the artifacts to some extent. Table 5 summarizes some state-of-art SR-based multi-modality image fusion methods.

Table 5
Some state-of-the-art SR-based multi-modality image fusion methods.

Methods	Model	Dictionary	Fusion rule
[19,31,65,67–69]	SR	Learned from a set of images [19,68,69] Learned from source images [31,65,67]	Maximum l_1-norm selection of representation coefficient vectors [19,69] Maximum l_2-norm selection of representation coefficient vectors [68] Maximum selection of (absolute) coefficient vector entries [65,67] Summing of representation coefficient vectors [31]
[29]	Group SR	Learned from a set of images	Maximum l_2-norm selection of representation coefficient vectors
[22]	Adaptive SR	Multiple dictionaries with different dominant directions learned from a set of images	Maximum l_1-norm selection of representation coefficient vectors
[27]	NNSR	Learned from source images	Maximum l_1-norm & sparseness selection of representation coefficient vectors
[30,44,45,70]	JSR	Learned from a set of images [44,70] Learned from source images [30,45]	Summing of representation coefficient vectors [44,70] Weighted averaging of representation coefficient vectors [30,45]

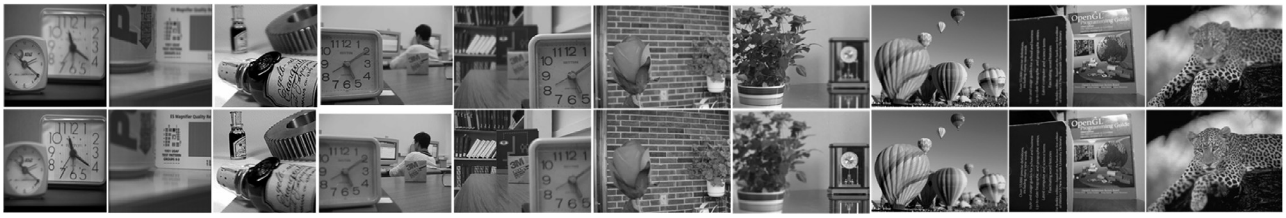

Fig. 11. 10 pairs of multi-focus test images. The top row contains 10 input images with the focus on the left part, and the bottom row contains the corresponding input images with the focus on the right part.

4.3. Applications to other fields of image fusion

In addition to the fusion of multi-focus images and multi-modality images, SR theory has also been exploited in other applications in image fusion. See for example these representative papers on remote sensing image fusion (also called *pan-sharpening*) [26,59,71–78] and multi-exposure [79] image fusion. SR models, dictionary construction (or learning) and fusion rules are also three key components in these SR-based remote image fusion methods. Further the traditional SR model is still the most popular in these methods. However, many of the dictionary construction and fusion rules employed in these methods are different from those in the fusion methods for multi-focus images and multi-modality images. Accordingly, many SR-based remote sensing image fusion methods are quite different from those used in SR-based multi-focus image fusion and multi-modality image fusion as discussed in the previous two sub-sections. This may be partially due to the fusion goal of pan-sharpening, which is to obtain a high spatial-resolution multispectral (HRM) image with the same spectral response as the multispectral (MS) sensor but the spatial-resolution of the panchromatic (PAN) sensor [59]. Considering these significant differences and the considerable length of the current paper, it seems best to employ a separate paper to properly describe these SR-based remote sensing image fusion approaches. This is such an important topic, and we have already started these investigations.

5. Experiments and analysis

As discussed in the previous sections, SR models, learned dictionaries and activity levels are three important issues in SR-based fusion methods. In this section, we will discuss the impacts of these three components on the fusion performance in the context of the previous two applications. For this purpose, we employ two sets of test images, as shown in Figs. 11 and 12. The two sets of test images contain 10 pairs of multi-focus images and 10 pairs of infrared and visible images, respectively.

In addition, we employ the mutual information (*MI*) [80], the gradient preservation fusion quality metric Q_G [81], and the structure similarity (SSIM) fusion quality metric Q_S [82] to evaluate different fusion methods on the basis of the amount of spatial information extraction. We also use two phase congruency-based fusion quality metrics Q_{ZP} [83] and Q_{PC} [10] to evaluate different fusion methods in terms of spatial consistency. For the metric Q_{ZP}, the zero-mean normalized cross-correlation (ZNCC) of the phase congruency maps between the fused image and input images is computed. While for the metric Q_{PC}, the phase congruency maps of the fused image and input images are computed separately, and the gradient preservation metric in [81] is also employed. Note that the size of local windows or blocks (if required) is set to 8 × 8 in this paper. Other parameters are set to their default values during the computation of these metrics. Larger values of these metrics indicate better performance of a fusion method.

In these experiments, the SR-based fusion methods are applied on a patch by patch basis. That is, the source images are first divided into many patches of the same size and then these patches are fused. The size of the patches is set to 8 × 8 as referring to the experimental results in [18]. Accordingly, the size of the dictionary atoms is also set to 8 × 8. In addition, in order to improve the robustness to mis-registration and reduce the spatial artifacts, a sliding window technology is employed, i.e., the patches overlap by one pixel.

5.1. SR Models

Next, the impact of different sparse representation models (listed in Table 6[9]) on the fusion performance will be discussed. In addition to those SR-based fusion methods in Table 6, some MST fusion methods, including NSCT, Curvelet and neighbor distance (ND) [84], are also compared, in which the simple "averaging" and "maximum absolute selection" fusion rules are employed to fuse the low-pass sub-band coefficients and the band-pass directional sub-band coefficients, respectively. Table 7 provides the scores of the different fusion methods on the two sets of test images, in which the average time T_a of different SR-based fusion methods are also provided. The experimental results in Table 7 indicate that the sparse representation model has a great effect on the fusion performance. As shown in Table 7, fusion performance varies sig-

[9] The dictionaries used in the models mentioned in Table 6 are learned from a database containing 24 high-resolution training images that are downloaded from http://r0k.us/graphics/kodak/.

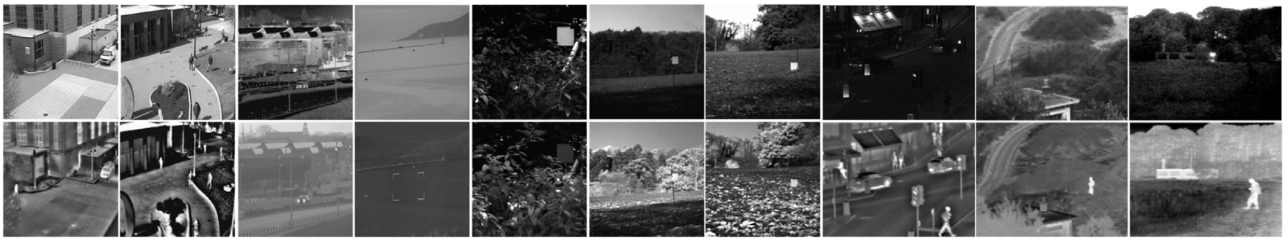

Fig. 12. 10 pairs of multi-modality test images. The top row contains 10 visible input images, and the bottom row contains the corresponding infrared input images.

Table 6
Different fusion methods with various SR models and their key parameters.

Model	Dictionary	Fusion rule
SR [15,18,19]	Learned from a set of images using K-SVD method [19], with 512 atoms	Maximum l_1-norm selection of representation coefficient vectors
ASR [22]	Multiple dictionaries [22] with different dominant directions learned from a set of images using K-SVD method, each of them with 512 atoms	
GSR [29]	Learned from a set of images using the patch-clustering method [31], with 489 atoms	
NNSR [27]	Learned from a set of images using the method in [42], with 512 atoms	
JSR [30,43,45]	Learned from a set of images using K-SVD method [19], with 512 atoms	
RSR [28]	Learned from a set of images using K-SVD method [19], with 512 atoms	Maximum l_2-norm of sparse reconstruction errors

Table 7
Performance of different SR models on the two sets of test images. Scores for all image pairs in each dataset are averaged.

Test images	Models	*MI*	Q_G	Q_S	Q_{ZP}	Q_{PC}	T_a (in seconds)
Multi-focus images	SR	4.1267	0.7584	0.5008	0.9533	0.6846	247.37
	ASR	4.0889	0.7548	0.4976	0.9444	0.6773	231.65
	GSR	4.6534	**0.7696**	0.5097	0.9587	**0.6940**	118.62
	NNSR	4.0504	0.7565	0.4994	0.9574	0.6615	7497.22
	JSR	4.6081	0.7666	0.5108	0.9565	0.6934	4222.82
	RSR	**4.8720**	0.7691	0.5024	**0.9681**	0.6916	29.71
	NSCT	3.9026	0.7473	**0.5134**	0.9467	0.6539	**1.53**
	Curvelet	3.8025	0.7273	0.4961	0.9281	0.6325	1.88
	ND	4.0648	0.7526	0.5036	0.9427	0.6619	2.45
Visible-infrared images	SR	2.3239	0.6192	0.4225	0.8340	0.4208	226.79
	ASR	2.2411	0.5966	0.4154	0.8284	0.4112	240.76
	GSR	**3.0451**	**0.6346**	0.4240	0.8658	0.4486	106.11
	NNSR	2.8963	0.6194	0.4152	**0.9025**	**0.4699**	7470.42
	JSR	2.4258	0.6178	0.4205	0.7815	0.3992	592.31
	RSR	2.7403	0.6335	**0.4320**	0.7936	0.4129	26.96
	NSCT	1.4811	0.5113	0.4432	0.7639	0.3374	**1.74**
	Curvelet	1.3105	0.5153	0.4268	0.7368	0.3121	2.00
	ND	1.3702	0.5978	0.4563	0.7807	0.3553	2.80

nificantly with the employed sparse representation model in an image fusion method. It also shows that the GSR performs the best among the six models considered here. In terms of most quality metrics, it achieves the highest scores for the fusion of multi-focus images as well as for the fusion of infrared and visible images. This may be due to the cluster structure sparsity prior employed in the GSR model. In addition to GSR, RSR and NNSR could also achieve satisfactory results when applied to multi-focus image fusion and multi-modality image fusion, respectively. However, for multi-modality image fusion, JSR could not achieve a satisfactory result as it did in [30]. This might be due to the employed dictionary KSVD-512 that was learned for SR rather than for JSR.

Table 7 also indicates that SR-based fusion methods generally perform better than the traditional MST fusion methods in information extraction and spatial consistency for the fusion of multi-focus images as well as for the fusion of multi-modality images, but it comes with the great cost of computational complexity. As shown in Table 7, those SR-based fusion methods, especially NNSR and JSR, are more computationally expensive than those traditional MST-based ones. This may be due to the greater time consumed in the sparse coding phase within these methods.

5.2. Dictionary construction

In this part, we will study the effect of the employed dictionary on the fusion performance. In all the experiments conducted, we employ the traditional SR model, and the maximum l1-norm as the fusion rule during the fusion process. Moreover, we test two kinds of over-complete dictionaries on the two sets of test images. The first is a 2-D over-complete DCT dictionary of size 512 (DCT-512, for short) [18]. The second includes four global trained dictionaries of size 128, 256, 512, and 1024. The four dictionaries (KSVD-128, KSVD-256, KSVD-512, and KSVD-1024, for short) are all learned from image samples using the iterative K-SVD algorithm [24]. The training data consist of 50,000 8 × 8 patches, randomly taken from the database mentioned in the previous Section 5.1. We also test three sets of adaptively trained dictionaries (denoted by D_{vi}-512,D_{ir}-512, and D_{joint}-512) on the infrared-visible test image set (i.e., the second set of test images). Each dictionary in the D_{vi}-512 set consists of 512 atoms and is learned from the corresponding visible input image in the second set of test images by using the iterative K-SVD algorithm. Similarly, each dictionary in the D_{ir}-512 set is learned from the corresponding infrared input

Table 8
Performance of different dictionaries on the two sets of test images. Scores for all image pairs in each dataset are averaged.

Test images	Dictionary	*MI*	Q_G	Q_S	Q_{ZP}	Q_{PC}
Multi-focus images	DCT-512	3.9947	0.7350	0.4732	0.8941	0.6443
	KSVD-128	3.8924	0.7439	0.4737	0.9012	0.6620
	KSVD-256	4.0344	0.7575	0.5003	0.9523	0.6826
	KSVD-512	**4.1267**	**0.7584**	**0.5008**	**0.9533**	**0.6846**
	KSVD-1024	4.0588	0.7532	0.4919	0.9321	0.6753
Visible-infrared images	DCT-512	2.3280	0.5892	0.3939	0.8195	0.4082
	KSVD-128	2.0021	0.5941	0.4009	0.7680	0.3858
	KSVD-256	2.2390	0.6179	0.4210	0.8287	0.4175
	KSVD-512	**2.3239**	**0.6192**	**0.4225**	0.8340	0.4208
	KSVD-1024	2.2528	0.6088	0.4158	0.8218	0.4124
	D_{vi}-512	2.3111	0.6121	0.4196	**0.8406**	**0.4218**
	D_{ir}-512	2.1703	0.6051	0.4126	0.8106	0.4059
	D_{joint}-512	2.2774	0.6109	0.4184	0.8357	0.4216

Table 9
Performance of different activity levels on the two sets of test images. Scores for all image pairs in each dataset are averaged.

Test images	Activity level	*MI*	Q_G	Q_S	Q_{ZP}	Q_{PC}
Multi-focus images	l_0-norm	**4.5006**	0.7098	0.4774	0.9761	0.6529
	l_1-norm	4.1267	**0.7584**	**0.5008**	**0.9533**	**0.6846**
	l_2-norm	4.0761	0.7557	0.5025	0.9473	0.6755
Visible-infrared images	l_0-norm	**2.9882**	0.5826	0.4037	**0.9679**	**0.5246**
	l_1-norm	2.3239	**0.6192**	**0.4225**	0.8340	0.4208
	l_2-norm	2.3390	0.6135	0.4217	0.8453	0.4242

image, and each dictionary in the D_{joint}-512 set is learned from the corresponding visible and infrared test images.

Table 8 provides the fusion scores of different dictionaries on the two sets of test images. According to Table 8: (1) As expected, the global learned dictionaries usually perform better than the fixed DCT dictionary. (2) The adaptively trained dictionaries in D_{vi}-512 and D_{joint}-512 sets, especially the ones in the former set, perform competitively with the global dictionary having the same number of atoms when applied to multi-modality image fusion. However, the dictionaries in the D_{ir}-512 set that are adaptively learned from the infrared input images do not perform better than the global learned dictionary and the ones in D_{vi}-512 and D_{joint}-512 sets. This may be due to the fact that fewer patches in the infrared images contain significant structures. As a result, the dictionaries in the D_{ir}-512 set have weak representation power and reduce the fusion performance. In contrast, the visible input images contain many more patches with significant structures. Correspondingly, the dictionaries in the D_{vi}-512 set seem to achieve better fusion performance. (3) It can also be argued that the number of dictionary atoms have a great impact on the fusion performance. As shown in Table 8, the dictionary KSVD-512 obtains the highest fusion performance among the four global dictionaries studied when applied to multi-focus image fusion as well as multi-modality image fusion. For the dictionary KSVD-128, the number of dictionary atoms seems too small, and some image patches (e.g., those with significant details) are not well represented. Therefore, the fusion performance is not comparable to the one obtained by using the dictionaries KSVD-256 and KSVD-512. However, if the number of dictionary atoms reaches certain values (e.g., 1024), there will be many more atoms with similar features in the dictionary. This may lead to the following fact. Those patches with similar features (e.g., focus information) within a multi-focus input image, even for those spatially-adjacent patches with greatly similar focus information, may be reconstructed by using different dictionary atoms and thus end up with different sparse representation coefficients. In the subsequent fusion process, these spatially-adjacent patches, even those with greatly similar focus information, are likely to be desynchronized in the fused image. As a result, spatial artifacts will be introduced to the fused image with the increase of the number of the dictionary atoms. And this can degrade the fusion performance to some extent. KSVD-1024 is one such example. In addition, this will also increase the computational complexity of a fusion method.

5.3. Activity levels

Thereafter, we discuss the impact of three activity level measures, l_0-norm, l_1-norm and l_2-norm of representation coefficients in (29), on the fusion performance. In this experiment, we employ the traditional SR model and the maximum-selecting fusion rule during the fusion process. The quantitative values obtained by the image fusion quality measures considered in Table 9 indicate that the l_1-norm of representation coefficients is a better choice among the three activity levels mentioned here. It achieves higher scores for the fusion of multi-focus images as well as for the fusion of multi-modality images, especially for the former.

6. Conclusion and discussion

SR-based image fusion methods have attracted much attention recently. Sparse representation models, dictionary learning, and fusion rules are three key components of in these techniques. In this paper, we have presented a thorough survey on the issues related to SR-based fusion methods. The following conclusions could be drawn accordingly.

For representation models, the traditional SR model is the most popular in image fusion. Extensions, such as ASR, GSR, NNSR, JSR, and RSR models, have also been applied to image fusion. Fusion performance varies with these models depending on the application. For example, GSR generally achieves better fusion performance when applied to multi-focus image fusion as well as infrared and visible image fusion. RSR and NNSR might also be a good choice for the fusion of multi-focus images and multi-modality images, respectively.

Regarding the dictionaries, the over-complete dictionaries with a fixed basis (e.g., a DCT basis) and those learned from a set of training images (*global trained dictionary*) or the input images themselves (*adaptively trained dictionary*) have been applied to image fusion. Generally, the learned dictionaries could achieve better fusion performance than those with a fixed basis. The number of atoms in a dictionary has a strong impact on the fusion performance. A compact dictionary with good representation capability is greatly desirable in image fusion for high fusion performance and computational efficiency. However, this is still a challenging problem in that area.

For fusion strategies, the l_0-norm, l_1-norm and l_2-norm of the representation coefficients or reconstruction errors are usually employed as the activity level. The maximum-selecting fusion rule is employed in most of the existing SR-based image fusion methods. Designing more sophisticated activity levels and fusion rules for SR-based image fusion methods presents an interesting research topic for the future.

There are some other issues related to the SR-based fusion methods that should be considered in future work. First, most of the SR-based image fusion algorithms have high computational complexity because of the increased time consumed during the sparse coding. This prevents SR-based methods from being used in the applications that demand real-time operation. Secondly, most of the current SR-based fusion methods are performed in a patch-based way. In order to improve the robustness to mis-registration while reducing the spatial artifacts, a sliding window technology is often employed. This results in the loss of detail information in the fused image and in the huge increase of computational complexity. Alternatively, the newly merged convolutional SR-based (CSR) fusion method [85] may be an interesting attempt to address such problems. Last, but not least, most of the SR-based fusion methods independently consider the local information from each image patch during the fusion process, including the sparse coding phase. In fact, strong correlations exist among each image patch and its spatially-adjacent neighbors. The fusion performance may be greatly improved if a local consistency prior is taken into account during the fusion process [28].

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 61104212, by Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2016JM6008), by the Fundamental Research Funds for the Central Universities under Grant No. JBZ170401, by the National Science Foundation under Grant No. ECCS-1405579, by the Royal Society Newton Mobility Grant IE150997, UK, and by the Australian Research Council Projects FT-130101457, DP-140102164, LP-150100671.

References

[1] Y. Liu, S. Liu, Z. Wang, Multi-focus image fusion with dense SIFT, Inf. Fusion 23 (2015) 139–155.
[2] S. Li, X. Kang, L. Fang, J. Hu, H. Yin, Pixel-level image fusion: a survey of the state of the art, Inf. Fusion 33 (2017) 100–112.
[3] Q. Zhang, Y. Wang, M.D. Levine, X. Yuan, L. Wang, Multisensor video fusion based on higher order singular value decomposition, Inf. Fusion 24 (2015) 54–71.
[4] S. Pertuz, D. Puig, M.A. Garcia, A. Fusiello, Generation of all-in-focus images by noise-robust selective fusion of limited depth-of-field images, IEEE Trans. Image Process. 22 (3) (2013) 1242–1251.
[5] A.P. James, B. Dasarathy, Medical image fusion: a survey of the state of the art, Inf. Fusion 19 (2014) 4–19.
[6] S. Li, X. Kang, J. Hu, Image fusion with guided filtering, IEEE Trans. Image Process. 22 (7) (2013) 2864–2875.
[7] S. Li, B. Yang, J. Hu, Performance comparison of different multi-resolution transforms for image fusion, Inf. Fusion 12 (2) (2011) 74–84.
[8] G. Pajares, J.M.D.L. Cruz, A wavelet-based image fusion tutorial, Pattern Recognit. 37 (9) (2004) 1855–1872.
[9] Z. Zhang, R.S. Blum, A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application, Proc. IEEE 87 (8) (1999) 1315–1326.
[10] Q. Zhang, Z. Ma, L. Wang, Multimodality image fusion by using both phase and magnitude information, Pattern Recognit. Lett. 34 (2) (2013) 185–193.
[11] Y. Liu, J. Jin, Q. Wang, Y. Shen, X. Dong, Region level based multi-focus image fusion using quaternion wavelet and normalized cut, Signal Process. 97 (7) (2014) 9–30.
[12] L. Guo, M. Dai, M. Zhu, Multifocus color image fusion based on quaternion curvelet transform, Opt. Express 20 (17) (2012) 18846–18860.
[13] L. Wang, B. Li, L. Tian, Multi-modal medical image fusion using the inter-scale and intra-scale dependencies between image shift-invariant shearlet coefficients, Inf. Fusion 19 (1) (2014) 20–28.
[14] K.P. Upla, M.V. Joshi, P.P. Gajjar, An edge preserving multiresolution fusion: use of contourlet transform and MRF prior, IEEE Trans. Geosci. Remote Sens. 53 (6) (2015) 3210–3220.
[15] J. Wright, A.Y. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[16] T. Guha, R.K. Ward, Learning sparse representations for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 34 (8) (2012) 1576–1588.
[17] X. Yuan, X. Liu, S. Yan, Visual classification with multitask joint sparse representation, IEEE Trans. Image Process. 21 (10) (2012) 4349–4360.
[18] B. Yang, S. Li, Multifocus image fusion and restoration with sparse representation, IEEE Trans. Instrum. Meas. 59 (4) (2010) 884–892.
[19] Y. Liu, S. Liu, Z. Wang, A general framework for image fusion based on multi--scale transform and sparse representation, Inf. Fusion 24 (2015) 147–164.
[20] Q. Zhang, X. Maldague, An adaptive fusion approach for infrared and visible images based on NSCT and compressed sensing, Infrared Phys. Technol. 74 (2016) 11–20.
[21] Q. Wei, J. Bioucas-Dias, N. Dobigeon, J.-Y. Tourneret, Hyperspectral and multispectral image fusion based on a sparse representation, IEEE Trans. Geosci. Remote Sens. 53 (7) (2015) 3658–3668.
[22] Y. Liu, Z. Wang, Simultaneous image fusion and denoising with adaptive sparse representation, IET Image Process. 9 (5) (2015) 347–357.
[23] M. Nejati, S. Samavi, S. Shirani, Multi-focus image fusion using dictionary-based sparse representation, Inf. Fusion 25 (2015) 72–84.
[24] 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.
[25] H. Yin, S. Li, L. Fang, Simultaneous image fusion and super-resolution using sparse representation, Inf. Fusion 14 (3) (2013) 229–240.
[26] M. Guo, H. Zhang, J. Li, L. Zhang, H. Shen, An online coupled dictionary learning approach for remote sensing image fusion, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7 (4) (2014) 1284–1294.
[27] J. Wang, J. Peng, X. Feng, G. He, J. Fan, Fusion method for infrared and visible images by using non-negative sparse representation, Infrared Phys. Technol. 67 (2014) 477–489.
[28] Q. Zhang, M.D. Levine, Robust multi-focus image fusion using multi-task sparse representation and spatial context, IEEE Trans. Image Process. 25 (5) (2016) 2045–2058.
[29] S. Li, H. Yin, L. Fang, Group-sparse representation with dictionary learning for medical image denoising and fusion, IEEE Trans. Biomed. Eng. 59 (12) (2012) 3450–3459.
[30] Q. Zhang, Y. Fu, H. Li, J. Zou, Dictionary learning method for joint sparse representation-based image fusion, Opt. Eng. 52 (5) (2013) 1–11.
[31] M. Kim, D.K. Han, H. Ko, Joint patch clustering-based dictionary learning for multimodal image fusion, Inf. Fusion 27 (2016) 198–214.
[32] G. Vivone, L. Alparone, J. Chanussot, M.D. Mura, A. Garzelli, G. Licciardi, R. Restaino, L. Wald, A critical comparison among pansharpening algorithms, IEEE Trans. Geosci. Remote Sens. 53 (5) (2015) 2565–2586.
[33] L. Loncan, L.B. de Almeida, J.M. Bioucas-Dias, X. Briottet, J. Chanussot, N. Dobigeon, S. Fabre, W. Liao, G.A. Licciardi, M. Simoes, et al., Hyperspectral pansharpening: a review, IEEE Geosci. Remote Sens. Mag. 3 (3) (2015) 27–46.
[34] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T.S. Huang, S. Yan, Sparse representation for computer vision and pattern recognition, Proc. IEEE 98 (6) (2010) 1031–1044.
[35] Z. Zhang, Y. Xu, J. Yang, X. Li, D. Zhang, A survey of sparse representation: algorithms and applications, IEEE Access 3 (2015) 490–530.
[36] S. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries, IEEE Trans. Signal Process. 41 (12) (1993) 3397–3415.
[37] A.M. Bruckstein, D.L. Donoho, M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, SIAM Rev. 51 (1) (2009) 34–81.
[38] J.A. Tropp, A.C. Gilbert, M.J. Strauss, Algorithms for simultaneous sparse approximation: part I: greedy pursuit, Signal Process. 86 (3) (2006) 572–588.
[39] E.J. Candes, X. Li, Y. Ma, J. Wright, Robust principal component analysis? J. ACM 58 (3) (2011) 1–37.
[40] D.L. Donoho, Y. Tsaig, Fast solution of ℓ_1 -norm minimization problems when the solution may be sparse, IEEE Trans. Inf. Theory 54 (11) (2008) 4789–4812.
[41] D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788–791.
[42] W. Dong, F. Fu, G. Shi, X. Cao, J. Wu, G. Li, X. Li, Hyperspectral image super-resolution via non-negative structured sparse representation, IEEE Trans. Image Process. 25 (5) (2016) 2337–2352.

[43] D. Baron, M.B. Wakin, M.F. Duarte, S. Sarvotham, R.G. Baraniuk, Distributed compressed sensing, IEEE Trans. Inf. Theory 52 (12) (2006) 5406–5425.
[44] H. Yin, S. Li, Multimodal image fusion with joint sparsity model, Opt. Eng. 50 (6) (2011) 1–10.
[45] N. Yu, T. Qiu, F. Bi, A. Wang, Image features extraction and fusion based on joint sparse representation, IEEE J. Sel. Top. Signal Process. 5 (5) (2011) 1074–1082.
[46] A. Majumdar, R.K. Ward, Fast group sparse classification, Can. J. Electr. Comput. Eng. 34 (4) (2009) 136–144.
[47] Z. Lin, R. Liu, Z. Su, Linearized alternating direction method with adaptive penalty for low-rank representation, in: Advances in Neural Information Processing Systems, 2011, pp. 612–620.
[48] Y. Zhang, Z. Jiang, L.S. Davis, Learning structured low-rank representations for image classification, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 676–683.
[49] C. Lang, G. Liu, J. Yu, S. Yan, Saliency detection by multitask sparsity pursuit, IEEE Trans. Image Process. 21 (3) (2012) 1327–1338.
[50] B. Yang, S. Li, Pixel-level image fusion with simultaneous orthogonal matching pursuit, Inf. Fusion 13 (1) (2012) 10–19.
[51] L. Landweber, An iteration formula for Fredholm integral equations of the first kind, Am. J. Math. 73 (3) (1951) 615–624.
[52] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, Cambridge, UK, 1985.
[53] J. Mairal, F. Bach, J. Ponce, G. Sapiro, A. Zisserman, Non-local sparse models for image restoration, in: Proceedings of IEEE International Conference on Computer Vision, 2009, pp. 2272–2279.
[54] P. Chatterjee, P. Milanfar, Clustering-based denoising with locally learned dictionaries, IEEE Trans. Image Process. 18 (7) (2009) 1438–1451.
[55] W. Dong, X. Li, L. Zhang, G. Shi, Sparsity-based image denoising via dictionary learning and structural clustering, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 457–464.
[56] J. Yang, J. Wright, T.S. Huang, Y. Ma, Image super-resolution via sparse representation, IEEE Trans. Image Process. 19 (11) (2010) 2861–2873.
[57] M.T. Iqbal, J. Chen, Unification of image fusion and super-resolution using jointly trained dictionaries and local information contents, IET Image Process. 6 (9) (2012) 1299–1310.
[58] K. Ren, F. Xu, Super-resolution images fusion via compressed sensing and low-rank matrix decomposition, Infrared Phys. Technol. 68 (2015) 61–68.
[59] X.X. Zhu, R. Bamler, A sparse image fusion algorithm with application to pan-sharpening, IEEE Trans. Geosci. Remote Sens. 51 (5) (2013) 2827–2836.
[60] R.Y. Ibrahim, J. Alirezaie, P. Babyn, Pixel level jointed sparse representation with RPCA image fusion algorithm, in: Proceedings of International Conference on Telecommunications and Signal Processing, 2015, pp. 592–595.
[61] Y. Zhang, Y. Chen, A new image-fusion technique based on blocked sparse representation, in: Proceedings of International Conference on Computer Science and Information Technology, 2014, pp. 53–60.
[62] Y. Yao, X. Xin, P. Guo, OMP or BP ? A comparison study of image fusion based on joint sparse representation, in: Proceedings of International Conference on Neural Information Processing, 2012, pp. 75–82.
[63] Y. Yao, P. Guo, X. Xin, Z. Jiang, Image fusion by hierarchical joint sparse representation, Cogn. Comput. 6 (3) (2014) 281–292.
[64] B. Yang, J. Luo, L. Guo, F. Cheng, Simultaneous image fusion and demosaicing via compressive sensing, Inf. Process. Lett. 116 (7) (2016) 447–454.
[65] X. Lu, B. Zhang, Y. Zhao, H. Liu, H. Pei, The infrared and visible image fusion algorithm based on target separation and sparse representation, Infrared Phys. Technol. 67 (2014) 397–407.
[66] Q. Zhang, L. Wang, H. Li, Z. Ma, Similarity-based multimodality image fusion with shiftable complex directional pyramid, Pattern Recognit. Lett. 32 (13) (2011) 1544–1553.
[67] M. Ding, L. Wei, B. Wang, Research on fusion method for infrared and visible images via compressive sensing, Infrared Phys. Technol. 57 (2013) 56–67.
[68] H. Yin, Sparse representation with learned multiscale dictionary for image fusion, Neurocomputing 148 (2015) 600–610.
[69] Y. Liu, S. Liu, Z. Wang, Medical image fusion by combining nonsubsampled contourlet transform and sparse representation, in: Chapter in Communications in Computer and Information Science, 484, 2014, pp. 372–381.
[70] B. Yang, S. Li, Visual attention guided image fusion with sparse representation, Optik 125 (17) (2014) 4881–4888.
[71] M.R. Vicinanza, R. Restaino, G. Vivone, M.D. Mura, J. Chanussot, A pansharpening method based on the sparse representation of injected details, IEEE Geosci. Remote Sens. Lett. 12 (1) (2015) 180–184.
[72] C. Jiang, H. Zhang, H. Shen, L. Zhang, Two-step sparse coding for the pan-sharpening of remote sensing images, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7 (5) (2014) 1792–1805.
[73] S. Li, H. Yin, L. Fang, Remote sensing image fusion via sparse representations over learned dictionaries, IEEE Trans. Geosci. Remote Sens. 51 (9) (2013) 4779–4789.
[74] H. Yin, Sparse representation based pansharpening with details injection model, Signal Process. 113 (C) (2015) 218–227.
[75] S. Li, B. Yang, A new pan-sharpening method using a compressed sensing technique, IEEE Trans. Geosci. Remote Sens. 49 (2) (2011) 738–746.
[76] X.X. Zhu, C. Grohnfeldt, R. Bamler, Exploiting joint sparsity for pansharpening: the J-SparseFI algorithm, IEEE Trans. Geosci. Remote Sens. 54 (5) (2016) 2664–2681.
[77] C. Han, H. Zhang, C. Gao, C. Jiang, N. Sang, L. Zhang, A remote sensing image fusion method based on the analysis sparse model, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9 (1) (2016) 1–15.
[78] Z.H. Nezhad, A. Karami, R. Heylen, P. Scheunders, Fusion of hyperspectral and multispectral images using spectral unmixing and sparse coding, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9 (6) (2016) 1–13.
[79] J. Wang, H. Liu, N. He, Exposure fusion based on sparse representation using approximate K-SVD, Neurocomputing 135 (2014) 145–154.
[80] G. Qu, D. Zhang, P.F. Yan, Information measure for performance of image fusion, Electron. Lett. 38 (7) (2002) 313–315.
[81] C.S. Xydeas, V. Petrovic, Objective image fusion performance measure, Electron. Lett. 36 (4) (2000) 308–309.
[82] G. Piella, H. Heijmans, A new quality metric for image fusion, in: Proceedings of IEEE International Conference on Image Processing, 2003, pp. 173–176.
[83] Z. Liu, D.S. Forsyth, R. Laganiere, A feature-based metric for the quantitative evaluation of pixel-level image fusion, Comput. Vis. Image Understanding 109 (1) (2008) 56–68.
[84] H. Zhao, Z. Shang, Y.Y. Tang, B. Fang, Multi-focus image fusion based on the neighbor distance, Pattern Recognit. 46 (3) (2013) 1002–1011.
[85] Y. Liu, X. Chen, R. Ward, Z.J. Wang, Image fusion with convolutional sparse representation, IEEE Signal Process. Lett. 23 (12) (2016) 1882–1886.

Qiang Zhang received the B.S. degree in automatic control, the M.S. degree in pattern recognition and intelligent systems, and the Ph.D. degree in circuit and system from Xidian University, China, in 2001,2004, and 2008, respectively. He was a Visiting Scholar with the Center for Intelligent Machines, McGill University, Canada. He is currently a professor with the Automatic Control Department, Xidian University, China. His current research interests include image fusion, computer vision and pattern recognition. He has published more than ten image-fusion papers, in which several of them have been well recognized in the field.

Yi Liu received the B. S. degree from Nanjing Institute of Technology, Nanjing, China, in 2012, and the M. S. degree from the Dalian University, Dalian, China, in 2015. He is currently working towards the Ph. D. degree in Control Theory and Control Engineering at Xidian University, Xian. His current research interests include computer vision and salient object detection.

Rick S. Blum received a B.S. in Electrical Engineering from the Pennsylvania State University in 1984 and his M.S. and Ph.D. in Electrical Engineering from the University of Pennsylvania in 1987 and 1991. From 1984 to 1991 he was a member of technical staff at General Electric Aerospace in Valley Forge, Pennsylvania and he graduated from GEs Advanced Course in Engineering. Since 1991, he has been with the Electrical and Computer Engineering Department at Lehigh University in Bethlehem, Pennsylvania where he is currently a professor and holds the Robert W. Wieseman Endowed Professorship in Electrical Engineering. His research interests include signal processing for smart grid, communications, sensor networking, radar, image fusion and sensor processing. He was on the editorial board for the Journal of Advances in Information Fusion of the International Society of Information Fusion. He was an associate editor for IEEE Transactions on Signal Processing and for IEEE Communications Letters. He has edited special issues for IEEE Transactions on Signal Processing, IEEE Journal of Selected Topics in Signal Processing and IEEE Journal on Selected Areas in Communications. He was a member of the SAM Technical Committee (TC) of the IEEE Signal Processing Society. He was a member of the Signal Processing for Communications TC of the IEEE Signal Processing Society and was a member of the Communications Theory TC of the IEEE Communication Society. He was on the awards Committee of the IEEE Communication Society. Dr. Blum is a Fellow of the IEEE, a former IEEE Signal Processing Society Distinguished Lecturer, anIEEE Third Millennium Medal winner, a member of EtaKappa Nuand SigmaXi,and holds several patents. He was awarded an ONR Young Investigator Award and an NSF Research Initiation Award. His IEEE Fellow Citation “for scientific contributions to detection, data fusion and signal processing with multiple sensors” acknowledges contributions to the fields of sensor processing and sensor networking.

Jungong Han is working in Lancaster University, UK. Previously, he was with the Northumbria University (2015–2017), UK, was with Philips CI (2012–2015), a research staff (2010–2012) with the Centre for Mathematics and Computer Science, and a researcher (2005–2010) with the Technical University of Eindhoven in Netherlands. Dr. Hans research interests include multimodality data fusion, computer vision, and artificial intelligence. He has written and co-authored over 100 papers, in which one first-authored paper has been cited, up to date, for more than 650 times. He is an associate editor of Elsevier Neurocomputing and Springer Multimedia Tools and Applications.

Dacheng Tao is Professor of Computer Science and ARC Future Fellow in the School of Information Technologies and the Faculty of Engineering and Information Technologies, and the Inaugural Director of the UBTech Sydney Artificial Intelligence Institute, at The University of Sydney. He mainly applies statistics and mathematics to Artificial Intelligence and Data Science. His research interests spread across computer vision, data science, image processing, machine learning, and video surveillance. His research results have expounded in one monograph and 500+ publications at prestigious journals and prominent conferences, such as IEEE T-PAMI, T-NNLS, T-IP, JMLR, IJCV, NIPS, CIKM, ICML, CVPR, ICCV, ECCV, AISTATS, ICDM; and ACM SIGKDD, with several best paper awards, such as the best theory/algorithm paper runner up award in IEEE ICDM’07, the best student paper award in IEEE ICDM’13, the 2014 ICDM 10-year highest-impact paper award, and the 2017 IEEE Signal Processing Society Best Paper Award. He received the 2015 Australian Scopus-Eureka Prize, the 2015 ACS Gold Disruptor Award and the 2015 UTS Vice-Chancellor’s Medal for Exceptional Research. He is a Fellow of the IEEE, OSA, IAPR and SPIE.