Multi-focus image fusion evaluation based on jointly sparse representation and atom focus measure

Yanxiang Hu  |  Bo Zhang  |  Juntong Zhang  |  Qian Gao

Abstract

Multi-focus image fusion (MFIF) tries to combine images with different in-focus regions and get a composite image that is in focus everywhere. Although many new MFIF algorithms based on various new representation models have been proposed in recent years, the performance evaluation of MFIF is still a challenging issue. In this study, a novel MFIF objective evaluation metric based on jointly sparse representation and atom focus measure is proposed. It not only provides a more reliable alternative for MFIF quality measurement but also supplies a unique MFIF performance analysis method at the same time. In the measurement, the sources and their fusion results are decomposed jointly sparse with an over-complete learning dictionary to extract the atom remnants of the source images. Meanwhile, in order to emphasise the fusion effect of in-focus atoms, the sum-modified-Laplacian model is used to measure the atom focus degree. Then, the atom remnants weighted by their focus measures are used to measure MFIF quality. In the experiments, nine recently proposed fusion algorithms were tested to contrast the proposed metrics with other four widely used objective metrics. The experimental results demonstrated the rationality and accuracy of our method. Moreover, it was also proved quantificationally that the fusion degree of atoms is directly related to their in-focus degree, the high in-focus degree atoms usually have poor fusion effect.

1 | INTRODUCTION

Multi-focus image fusion (MFIF) aims to generate an all-in-focus image by integrating the images with local un-focus regions or different focal settings artificially. It provides more image details for subjective observations or machine vision processing. By now, MFIF has been applied in many imaging applications, such as machine vision, consumer electronics, medical imaging and remote sensing and so forth, and it is also an active topic in information fusion research [1].

As new signal processing models and methods are employed in image fusion, the pixel-level fusion algorithms can now be divided into four categories roughly: Spatial domain-, transform domain-, sparse representation (SR)- and neural network-based methods [2, 3]. Compared with the widely used multi-scale transform-based methods, learning dictionary-based SR has led to state-of-the-art fusion performance and presented persuasive potentials [3].

In MFIF, the evaluation of focus and the selection of combinational weights are the two core issues. Many local focus degree measures, such as local energy of Laplacian, spatial frequency, the sum of high-frequency coefficients in various transform domains, L0 or L1 norms of SR coefficient vectors and other more complex indices have been used [2, 3]. Based on focus assessment, weight score maps or decision maps of source images are constructed and then are used to guide the fusion process.

Fusion quality evaluation is another important topic in image fusion research. Although many objective evaluation metrics have been proposed, the selection of persuasive and pertinent metrics is still a question without a definitive answer [2–5]. The evaluation methods face the difficulty of how to evaluate the integration degree of salient information while identifying the authenticity of the fused results. Fortunately, the continuous evolution of signal representation and analysis methods provides new means and observation views for this problem. In
this study, we report a novel evaluation and analysis method for MFIF based on jointly SR (JSR) and atom focus measure. The contributions of this method include:

1. Assessing the fusion quality in SR domain. Because the atoms in SR-learning dictionaries are the building blocks of source images, so similarity comparisons at atom level are more reasonable than man-made low-level intensity feature comparisons. Specially, we take atom in-focus degree into account to emphasise the fusion effect of in-focus atoms. This makes our metric more task-specific.

2. Surpassing the existing metrics, the proposed method can analyse the atom-level characteristics of MFIF algorithms besides quality evaluation. This advantage makes it valuable in further algorithm analysis and improvement. Our experimental results demonstrated quantitatively that the in-focus atoms usually have relative poor fusion effect.

The study is organised as follows. SR, JSR and their applications in image fusion are introduced briefly in Section 2. In Section 3, the fusion quality evaluation metrics are discussed. The principle and computational process of our method are explained in depth in Section 4. In Section 5, the experimental results are reported and discussed in detail. The conclusions are reported in Section 6.

2 | SR AND JSR

2.1 | SR

SR is a new kind of signal representation method originating from compressed sensing and has been widely used in many computer vision tasks [6, 7].

In SR, a signal is represented as a linear combination of as few as possible dictionary atoms. Suppose $Y$ is an $n$-dimensional signal and $D \in \mathbb{R}^{m \times n}$ ($M > n$) is an over-complete dictionary, then $Y$ can be formulated as

$$\arg \min_{X} \|X\|_{0} \text{ s.t. } Y = DX$$

(1)

where $X \in \mathbb{R}^{M}$ denotes the SR coefficient vector and $\|X\|_{0}$ denotes L0 norm of $X$. Equation (1) seeks to construct signal $Y$ by using as few as possible atoms, so $X$ is sparse. The over-completeness guarantees SR can decompose and reconstruct the signals accurately, while the sparseness constraint can make the features contained in high-dimensional space approximate it better with a linear combination of low-dimensional sub-spaces. These merits are of especial value in visual information processing. The optimisation of Equation (1) is an Non-deterministic Polynomial hard problem. The existing solving approaches include greedy pursuits, convex relaxation and other approaches [7]. Among these approaches, orthogonal matching pursuit (OMP) has been used widely due to its high efficacy [7]. Besides solving approaches, the construction method of the over-complete dictionary is also a key issue in SR. It has been proved that learning dictionaries generated from trained samples outperform fixed basis dictionaries [3].

2.2 | JSR

As an extension of SR, JSR tries to describe a group of signals belonging to the same ensembles with the SR model [8]. These signals observe the same scene but possess different modalities or generating parameters. So they can be represented as the sum of a common component and respective innovation components together. Suppose, there are $K$ such different source signals $Y_{k} \in \mathbb{R}^{n}$ ($k = 1, \ldots, K$), $D \in \mathbb{R}^{K \times M}$ ($n < M$) is an over-complete dictionary. According to the concept of JSR, signal $Y_{k}$ can be represented as

$$Y_{k} = Y_{k}^{C} + Y_{k}^{U} = DX_{k}^{C} + DX_{k}^{U}$$

(2)

where $Y_{k}^{C}$ denotes the common component, and $Y_{k}^{U}$ denotes the innovation component of $Y_{k}$. $X_{k}^{C}$ and $X_{k}^{U} \in \mathbb{R}^{M}$ are the SR coefficient vectors of the common and innovation components, respectively. Matrix JSR of ensemble $Y$ can be given as

$$Y = \begin{bmatrix} Y_{1} \\ \vdots \\ Y_{K} \end{bmatrix} \in \mathbb{R}^{K \times n}, \quad D = \begin{bmatrix} D_{1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & D_{K} \end{bmatrix} \in \mathbb{R}^{K \times (K+1)M}, \quad X = \begin{bmatrix} X_{1}^{C} \\ \vdots \\ X_{K}^{C} \\ X_{1}^{U} \\ \vdots \\ X_{K}^{U} \end{bmatrix} \in \mathbb{R}^{(K+1)M \times 2}.$$  

(3)

where $Y$ denotes the joint column vector matrix, $D$ is the joint dictionary and $X$ is the joint coefficient matrix. In $D_{k} \in \mathbb{R}^{M \times 2}$ denotes all -0 sub-matrix.

2.3 | SR-based image fusion

The main steps of SR-based image fusion methods include: (1) The source images are separated into overlapping patches with sliding window technique and then each patch is rearranged as a column vector; (b) all patch vectors are SR-decomposed; (c) the SR coefficient vectors of the corresponding patches are combined according to the fusion rules to generate the fused SR coefficient vectors; (d) the fused results are reconstructed with the fused SR coefficient vectors and over-complete dictionaries [3].

In [9], Yang et al. introduced an MFIF method based on the non-fixed-base SR model and non-subsampled shearlet transform (NSST). The sources were decomposed with NSST at first. A non-fixed-base dictionary was trained iteratively and used to fuse the low-frequency components. At the same time, a type-2 fuzzy logic scheme was designed to fuse NSST high-frequency coefficients. In [10], a robust fusion method based on SR and online dictionary learning was proposed. Several key issues including block SR, restoration methods, feature extraction, dictionary learning method and fusion rules were investigated in detail. Zhang et al. presented a robust SR model to deal with non-Gaussian noise or sparse outliers [11]. Besides
the sparsity constraints, the conventional least-squared reconstruction error is replaced by a sparse reconstruction error. He presented a combinatorial method to implement multi-scale SR for image fusion [12]. Specifically, they learned an SR dictionary in non-subsampled contourlet transform (NSCT) domain to improve the description capability. In addition, Liu et al. also presented a general framework to make full use of the advantages of multi-scale transform (MST) and SR [13]—the sources are MST decomposed into low-pass and high-pass subbands first, and then the low-pass subbands are fused with an SR-based approach while the high-pass subbands are fused using a simple maximum selection rule. Nejati et al. employed SR in the generation of decision map for MFIF [14]. An adaptive-learning dictionary is trained first from focus patches and the focus features are extracted. Then the correlations between the learning dictionary and the sources are used to produce a pixel-level decision map. After optimisation, this map is used to reconstruct the fusion result. Recently, Li et al. reported a discriminative low-rank sparse dictionary learning-based method to implement image fusion, denoising and enhancement simultaneously [15]. In order to promote the discriminative capability of the learned dictionary, they integrated low-rank and sparse regularisation terms into the dictionary learning process. Furthermore, a weighted nuclear norm and a sparse constraint were imposed on the sparse components to eliminate noises and preserve details. More comprehensive reviews can be found in [2, 3, 16].

3 | COMMENTS ON FUSION QUALITY EVALUATION

In view of their good practicality and high efficiency, objective evaluation metrics with no references have been widely used in image fusion quality measurement. Some simple metrics measure the self-information contents of fusion results directly, such as entropy, average gradient, standard deviations and so forth. Owing to the lack of similarity comparisons with the sources, these metrics cannot indicate the deviations or alias caused by the fusion methods. By now, information theory-based methods and local feature similarity comparison-based methods are the two most important categories [2–5].

Information theory-based methods, such as mutual information (MI) [17] and normalised MI (NMI) [18], measure the dependence degree between fusion results and source images by mutual entropy. Suppose two input images are A, B and their fused result is F, then MI is defined as

\[ MI(A, B, F) = IEA(f, a) + IFB(f, b) \]

\[ IEA(f, a) = \sum_{f,a} p_{F,A}(f,a) \log \frac{p_{F,A}(f,a)}{p_{A}(a)} \]

\[ IFB(f, b) = \sum_{f,b} p_{F,B}(f,b) \log \frac{p_{F,B}(f,b)}{p_{B}(b)} \]

where \( a \) and \( b \) denote the intensity values of a pair of corresponding pixels in \( A \) and \( B \), and \( f \) denotes their fusion result. 

\[ p_{A}(a) \] denotes the marginal probability distribution of \( a \) in \( A \), and \( p_{F,A}(f,a) \) is the joint probability distribution between \( f \) and \( a \) and so forth.

To overcome the potential drawback that the entropies of two source images are un-normalised, Hossny et al. suggested revising MI by employing NMI measure:

\[ M^{NMI}_{F} = 2 \left( \frac{IEA(f,a)}{H(F) + H(A)} + \frac{IFB(f,b)}{H(F) + H(B)} \right) \]

where \( H(F), H(A) \) and \( H(B) \) denotes the entropies of \( F, A \) and \( B \), respectively.

Local feature similarity-based methods measure the transfer amount of local features from sources to fusion results. Some elementary spatial features, such as edge, gradient, contrast and so forth, have been used. In [19], a gradient-based metric \( Q^{AB/F} \) was proposed. Sobel edge detector is used to extract the edge information in the sources and the fused results at first, and then the similarities between edge contents are taken as the fusion quality indicator. In [20], a more comprehensive metric, termed as Visual Information Fidelity for image Fusion (VIFF), was designed. This metric evaluates fusion quality with the aid of MST and block-wise visual information fidelity analysis. In addition, structural SIMilarity index (SSIM)-based metrics [21], and multi-scale similarity metric [22] have also been proposed. Table 1 outlines some objective evaluation metrics commonly used. More introductions can be found in [2–5].

Although obvious progress has been made in fusion quality assessment, the selection of suitable indices is still a difficult task. Due to the diversity of observation angles, researchers tend to believe that it is hard to say that a certain metric is always better than others. To contrast these metrics quantitatively, some deep investigations have been made. For example, Liu et al. employed six MST fusion algorithms to contrast 12 objective metrics [4]. In [5], Li et al. tested six MST-based fusion methods with different transform settings and evaluated their quality with several commonly used metrics. Their work demonstrated that different indices show obvious diversities.

Specifically, the following drawbacks prevent the existing metrics from more reasonable evaluation.

1. Similarity comparison-based metrics employ some local features, such as gradients and contrasts. These primary features have no direct relationships with the characteristics of fusion tasks or image formation.

2. All existing metrics try to conclude fusion quality with a holistic value without further analysis, so they cannot offer more useful clues for algorithm improvement.

In this study, we present a JSR-based fusion evaluation method for MFIF. As far as we know, this is the first SR-based method for MFIF quality evaluation. This method assesses MFIF algorithms from the point of view of transfer integrity of atoms. By viewing the source images and their fusion result as an ensemble about the same scene, we use JSR model to decompose the sources and fusion result into a common component and different private components. The innovations of
TABLE 1 The objective evaluation metrics for image fusion

| Categories         | Typical metrics                                                                 | Artefact sensitivity |
|--------------------|----------------------------------------------------------------------------------|----------------------|
| Self Information   | Average gradient, Standard deviation, entropy and so forth                       | No                   |
| Mutual Information | Mutual information (MI), normalised MI (NMI), $Q^{MFIF}$, structural SIMilarity index (SSIM), VIFF, $Q_1$ [23], $Q_2$ [23] and so forth | Yes                  |

FIGURE 1 The flowchart of the proposed method

source images can be regarded as the combination of un-fused atom remnants. Specific to MFIF, the in-focus atoms should occupy larger weights in quality evaluation, so we take atom focus degrees into account further. Then, the residual atom coefficients and the corresponding focus measures of atoms are combined to measure the performance of MFIF algorithms. Compared with the existing metrics, atom-level similarity comparisons not only have a more explicit visual interpretation but also make it feasible to conduct fusion effect analysis further.

4 THE PROPOSED METRIC

Figure 1 gives the flowchart of our method. The four steps of our method are decomposing the sources and the fusion result sparsely jointly, computing the focus measure of atoms (FMA), computing the residual coefficient ratios and generating the final quality metric.

4.1 JSR decomposition

Without loss of generality and for clarity, suppose two source images are $A$ and $B$, their fused result is $F$. The size of these images is $M*N$. When the size of the patch is 8*8 and the sliding step length is 1, each image is divided into $L = (M – 8 + 1) \times (N – 8 + 1)$ patches. Each patch is then rearranged as a 64*1 vector in column-major order. After vectorisation, each image is transformed into a new matrix with the size of 64*1. These reshaped matrices are denoted as $YA$, $YB$ and $YF$ below. Next, $YF$ is JSR decomposed with $YA$ and $YB$, respectively, according to Equation (3):

\[
\begin{bmatrix}
YF \\
YA
\end{bmatrix} = \begin{bmatrix}
D1D0 \\
D1 0 D
\end{bmatrix} \begin{bmatrix}
X_{FA}^C \\
X_{FA}^U \\
X_{FA}^{U, AF}
\end{bmatrix}, \begin{bmatrix}
YF \\
YB
\end{bmatrix} = \begin{bmatrix}
D1D0 \\
D1 0 D
\end{bmatrix} \begin{bmatrix}
X_{FB}^C \\
X_{FB}^U \\
X_{FB}^{U, BF}
\end{bmatrix}
\]

(6)

where $D \in R^{64 \times K}$ is the over-complete learning dictionary used, and $K$ is the number of atoms contained in $D$. Sub-dictionary $D1$ is the column normalisation version of $D$. $X_{FA}^C$ and $X_{FB}^C \in R^{K \times L}$ are the SR coefficient matrices of common components between $YF/YA$ and $YF/YB$, respectively. $X_{FA}^U \in R^{K \times L}$ denotes the SR coefficient matrix of innovation component of $YF$ compared with $YA$. $X_{FB}^U \in R^{K \times L}$ denotes the SR coefficient matrix of innovation component of $YA$ compared with $YF$. The meanings of $X_{FB}^{U, BF}$ and $X_{FA}^{U, AF}$ are similar to the $X_{FA}^U$ and $X_{FB}^U$.

From Equation (6), it can be known that when atoms contained in $A$ or $B$ are fused completely into $F$, $X_{AF}^U$ and $X_{BF}^U$ should be all –0. So we call $X_{AF}^U$ and $X_{BF}^U$ as residual coefficient matrices to underline their meanings below.

As an illustration, the JSR decomposition results of ‘clock’ multi-focus images are shown in Figure 2. The source 1, source 2 and the fusion result fused with the simple average algorithm are shown in Figure 2(a). This fusion result was decomposed sparsely jointly with source 1 using the dictionary shown in Figure 3. The resulting JSR coefficient matrices were reconstructed as ‘component images’ and shown in Figure 2(b), including the common image, the residual image of the fusion result and source 1.
residual image of source 1 (from left to right). For clarity, the residual images shown in Figure 2(b) have been normalised to intensity range of the common image shown in Figure 2(a), and the two residual images shown in Figure 2(b) have been enlarged 50 times.

From this example, we can note the evidence that the residual information remained in source 1. This indicates that the fusion degree of source 1 is not ideal. Moreover, we can note that the remnants mainly appear near the edges and contours. This indicates intuitively that the high-focus atoms seem to have poor fusion effects.

4.2 Residual coefficient ratio

In Equation (6), each column of $X_{AF}^U$ and $X_{BF}^U$ constructs a 'residual patch' and each row contains all residual coefficients of an atom in all image patches. To compute the residual coefficients of $i$th atom, we define residual coefficient vector (RCV) $RCV \in \mathbb{R}^{K \times I}$ as

$$RCV(i) = \sum_{j=1}^{I} \left( |X_{AF}^U(i,j)| + |X_{BF}^U(i,j)| \right), i = (1, \ldots, K) \quad (7)$$

where $RCV(i)$ is the sum of absolute values of residual coefficients of $i$-th atom in all patches. Based on RCV, we define residual coefficient ratio (RCR) $RCR \in \mathbb{R}^{K \times I}$ as

$$RCR(i) = RCV(i) / \sum_{i=1}^{K} RCV(i), \quad i = (1, \ldots, K) \quad (8)$$

where $RCR(i)$ denotes the proportion of $i$-th atom remnant in the total atom remnants. The following experiments will demonstrate that the in-focus atoms usually have higher RCR compared with the smooth (un-focus) atoms.

4.3 Focus measure of atoms

In order to make our metric more task-specific, atom in-focus degree is taken into account in our method. In [24], Huang et al. compared some image focus measure metrics, such as variance, energy of image gradient, energy of Laplacian, sum-modified-Laplacian (SML) and spatial frequency and so forth. Their study showed that SML can provide better accuracy. In this study, we define FMA as the average of the sum of SML of non-edge pixels:

$$FMA = \frac{\sum_{j=2}^{7} \sum_{i=2}^{7} \left( |p(i,j)_{SML}| + |p(i,j)_{SML}| \right)}{36} \quad (9)$$

where $p(x_jy_j)_{SML}$ and $p(x_jy_j)_{SML}$ denote the second-class deviations of pixel $p(i,j)$. Here, we suppose the size of the atom is 8*8 and the coordinate origin of atoms is (1,1).

4.4 The proposed MFIF evaluation metric

The proposed metric $Q_{\text{JSR-FMA}}$ is defined as the average of the product of all atoms’ RCRs and their FMAs:

$$Q_{\text{JSR-FMA}} = \frac{\sum_{i=1}^{K} RCR(i) \times FMA(i)}{(2 \times L)} \quad (10)$$

where $K$ is the number of dictionary atoms and $L$ is the number of patches within one image. This metric is determined by the atom’s residual ratios and focus degrees and emphasises the fusion effect of in-focus atoms and the weighted residual coefficient of atoms by its focus measure. The smaller $Q_{\text{JSR-FMA}}$ indicates better fusion quality.

5 EXPERIMENTS AND DISCUSSION

5.1 Experimental settings

To validate our method, a large amount of MFIF experiments were carried out and the experimental results were measured with several recognised objective evaluation metrics with our metrics.

Four commonly used fusion evaluation metrics, including MI [17], NMI [18], $Q_{1B/F}$ [19] and VIFF [20] were employed to contrast with our metrics. Nine recently developed MFIF algorithms are evaluated with the above-mentioned metrics. These algorithms include: (1) Kumar's cross-bilateral filter(CBF)-based MFIF algorithm [25]; (2) Liu’s convolutional neural network (CNN)-based MFIF algorithm [26]; (3) Liu’s convolutional sparse representation(CSR)-based general fusion algorithm [27]; (4) Kumar’s discrete cosine harmonic wavelet transform (DCHWT)-based multi-focus and multi-spectral algorithm [28]; (5) Liu’s dense Scale-Invariant Feature Transform (DSIFT)-based MFIF algorithm [29]; (6) Li’s guided filtering fusion (GFF)-based general fusion algorithm [30]; (7) Li’s image matting fusion (Image Fusion with Matting, IFM)-based MFIF algorithm [31]; (8) Liu’s NSCT-SR general fusion algorithm [13]; (9) Nejati’s SR-based MFIF algorithm (Sparse Representation-based Classification for Fusion, SRCF) [14]. The open-source codes of these algorithms were
run with their default parameters. For the consistency of comparison, a global learning dictionary of 256 8*8 atoms [13] was employed in all our tests as shown in Figure 3. This dictionary was learned from 100,000 trained samples.

Twenty pairs of grayscale multi-focus source images were used for the test as shown in Figure 4 (numbered from left to right and from top to down). Here, only the grayscale sources were tested since most of the tested algorithms do not support colour MFIF. But as in [14], the proposed metric can evaluate colour MFIF by measuring the fusion quality of R, G and B channels, respectively, and then generating the final index. The sizes of these test images range from 160*160 (11th) to 944*736 (16th).

5.2 Experimental results and analysis

In this section, the following experiments are presented: (1) Twenty pairs of MFIF source images were fused with nine fusion algorithms and then measured with five objective evaluation metrics; (2) two groups of typical local regions of fusion results are shown for visual observations and subjective-objective measure consistency comparisons; (3) the characteristics of atom remnants in the above MFIF tests are analysed and discussed in detail.

5.2.1 Objective quality evaluation and comparison

Due to the limit of length, here we only report the final statistical ranking of nine fusion algorithms according to all 20 groups of fusion tests. The averages of standard deviation (ASDs) of algorithm rankings under each metric are also listed in the last row of Table 2.

From Table 2, the following conclusions can be found:

1. The proposed metric shows good consistency with MI, NMI, and \( Q^{1B/F} \), while VIFF shows an interesting difference with others—the nine algorithms get similar ranking except for NSCT-SR with gets the best place.
2. According to ASDs, all metrics show a high degree of consistency in all experiments, that is, the algorithm ranking in each experiment is highly consistent with the final overall ranking. NMI and our metric show the best evaluation stability.
3. Among nine algorithms under-tested, SRCF presents the best quality and DCHWT gets the worst ranking.

5.2.2 Subjective visual comparison

To compare the fusion results visually, two groups of typical local regions selected from 7th and 12th test are zoomed in and shown in Figures 5 and 6.

In Figure 5, a region with complex intensity changes and edge changes was picked to contrast the fusion effect. We can note that near the fence twisting, a certain degree of halo artefacts appears in all fusion results. The small dot near the twisting gets obviously attenuated in several algorithms and even been loosed in CSR. Some obvious artefacts in different fusion results are marked with red arrows. From the overall focus comparison, SRCF, DSIFT, CNN and GFF seem to get better edge contrast.

In Figure 6, a small region containing a long and sharp edge was selected to observe the focus region's distinguishing...
TABLE 2  The comprehensive ranking results of nine algorithms

| Metric       | MI    | NMI   | $Q_{AB/F}$ | VIFF | $Q_{SRF-FMA}$ |
|--------------|-------|-------|------------|------|--------------|
| Ranking      | (1) SRCF | (1) SRCF | (1) SRCF | (1) NSCT-SR | (1) SRCF |
|              | (2) DSIFT | (2) DSIFT | (2) CNN | (2) SRCF | (2) DSIFT |
|              | (3) CNN | (3) CNN | (3) DSIFT | (3) CNN | (3) CNN |
|              | (4) IFM | (4) IFM | (4) GFF | (4) DSIFT | (4) IFM |
|              | (5) GFF | (5) GFF | (5) IFM | (5) GFF | (5) GFF |
|              | (6) CSR | (6) CSR | (6) CBF | (6) IFM | (6) NSCT-SR |
|              | (7) NSCT-SR | (7) NSCT-SR | (7) CSR | (7) CSR | (7) CSR |
|              | (8) CBF | (8) NSCT-SR | (8) CSR | (8) CBF | (8) CBF |
|              | (9) DCHWT | (9) DCHWT | (9) DCHWT | (9) DCHWT | (9) DCHWT |
| Averages of standard deviation | 1.1842 | 0.5872 | 1.3811 | 1.7038 | 0.6726 |

FIGURE 5  The local region comparisons of 7th group of fusion results (a) two source images, (b) two observed regions, (c) nine fusion regions (from left and right and from top to down: Cross-bilateral filter (CBF), convolutional neural network (CNN), convolutional sparse representation (CSR), DCHWT, DSIFT, guided filtering fusion (GFF), IFM, non-subsampled contourlet transform sparse representation (NSCT-SR) and SRCF)

capability of fusion algorithms. We can note that SRCF, DSIFT, CSR and IFM get a better in-focus edge, but at the same time, obvious speckles or jagged blocks appear in the results of CSR, DSIFT and IFM. In the other five algorithms, the focus degrees of the long edge are obviously poor.

Besides intuitive observation, a subjective evaluation was also carried out on the fusion results as shown in Figures 5 and 6. Two teachers and four graduate students engaging in image research were selected as observers. In the tests, two sources and nine fusion results were shown on an HD LCD monitor, and the viewers were told to keep a fixed distance and score the fusion algorithms. A group of nine fusion images were given the scores 1–9, denoting the best to the worst. The sort results of all viewers were averaged and regarded as the final subjective reference. This subjective comparison indicates SRCF has the best fusion effect.

Moreover, the subjective-objective evaluation consistency was also investigated. Five groups of objective

FIGURE 6  The local region comparisons of 12th group of fusion results (a) two source images, (b) two source regions, (c) nine fusion regions (from left and right and from top to down: CBF, CNN, CSR, DCHWT, DSIFT, GFF, IFM, NSCTSR and SRCF)
FIGURE 7 Spearman correlation coefficients of subjective and objective evaluation

FIGURE 8 Residual coefficient ratio (RCR) distributions of 6th group of test of SRCF and CNN algorithms (a) SRCF, (b) CNN

comparison rankings of each group of fusion results were computed. Then the Spearman correlation coefficients (SROCC) of five objective metrics compared with subjective rankings were computed and shown in Figure 7.

It is obvious in Figure 7 that the five objective metrics all have a positive correlation with subjective evaluation and our metric gets the best score.

5.2.3 Analyses of atom fusion effect

In this subsection, we report the detailed atom-level analysis results. Some latent remarkable laws were revealed through our experiments.

1. At first, atom’s statistical rankings according to their RCRs are reported;
2. next, we analyse the regulations of atom content ratios in the sources;
3. finally, we investigate the relationships between RCR, atom content ratios and FMA.

According to Equation (7), we can get the RCRs of all atoms. These indexes indicate intuitively the size of remnants of atoms. Figure 8 shows the RCR distributions of SRCF and CNN results in the 6th group test. For clarity, the RCRs of 256 atoms have been sorted according to their absolute values. Because the number of atoms in the dictionary is 256, so the average RCR is 1/256 = 0.0039.

From Figure 8, we can note that:

1. Two RCR distributions have the same shapes. This conclusion also holds in all our experiments. This demonstrated that the atom RCR distributions are less affected by the fusion algorithms.

2. About 2/3 atom’s RCRs are less than the average RCR and a small part of atoms have very large RCR. It was found that the max RCRs in all experiments are 0.0146-0.0148 and are almost unaffected by the fusion algorithm too. Table 3 lists the top 10 atoms with the largest RCR in the fusion results of SRCF, DSIFT and CNN in the 6th group of test.

In Table 3, the atoms with the largest RCR are almost the same in various experiments. To explain this regular phenomenon, we compute the content ratios of atoms (CRA) in the source images: Sparsely decomposing the two source images A and B with the dictionary used, respectively:

\[ A = DX_A \text{ and } B = DX_B \quad (11) \]

and then computing the ratio of SR coefficient of each atom in the SR coefficient matrices \( X_A \) and \( X_B \):

\[ X = |X_A| + |X_B|, \quad CRA(i) = \frac{\sum_{j=1}^{L} X(i,j)}{\sum_{i=1}^{K} \sum_{j=1}^{L} X(i,j)} \quad (12) \]

where the definitions of \( K \) and \( L \) are as before.

In the comparisons of 9*20 groups of MFIF sources, it was found that just like RCR distributions, the CRA distributions also have a high degree of similarity. Table 4 shows the top 10 atoms with the largest CRA in 2nd, 7th and 12th group of source images.

It can be noted that the atom sequences listed in Table 4 have high similarity and they also have high consistency with the sorting results in Table 3. This demonstrates that the RCRs of atoms have a direct proportional relationship with their CRAs. Figure 9 plots the CRAs of all atoms (blue) of 6th group of source images. For easy comparison, CRAs have been sorted by their size. The corresponding RCRs (red) resulted from SRCF and FMAs (green) were also plotted together for contrast.

| Atom serial number |
|--------------------|
| SRCF               |
| 206, 23, 174, 177, 240, 77, 165, 163, 24, 152 |
| DSIFT              |
| 23, 206, 174, 240, 177, 77, 165, 163, 24, 152 |
| CNN                |
| 43, 23, 240, 174, 177, 77, 165, 163, 24, 152 |

| Atom serial number |
|--------------------|
| 2nd                |
| 27, 177, 206, 255, 43, 163, 77, 165, 24, 152 |
| 7th                |
| 206, 164, 255, 27, 77, 43, 165, 163, 24, 152 |
| 12th               |
| 237, 206,164, 27, 43, 255, 77, 165, 152 |

TABLE 3 Top 10 atoms with the largest RCR in 6th group of test

TABLE 4 Top 10 atoms with the largest CRA in 2nd, 7th and 12th groups of source images
FIGURE 9 Sorted focus measure of atoms (FMAs), content ratios of atoms (CRAs) of the 6th group of source images and RCRs resulted from SRCF

FIGURE 10 The ratio curve between RCR and CRA

In Figure 9, the X-axis denotes sorted atom number according to CRA, and Y-axis denotes the values of CRA/FMA/RCR of these atoms. It can be seen from this figure that:

1. The atoms with large CRA usually have high FMA, that is, in-focus atoms tend to own larger proportion in source images;
2. RCR and CRA curves show high similarity. This means the atoms owning large proportion in source images still have large remnants after fusion.

On the other hand, from the point of view of fusion quality comparison, it is valuable to make a further comparison between RCR and CRA. Figure 10 shows the ratio curve between RCR and CRA curves shown in Figure 9. For clarity, this curve has been smoothed slightly to remove small burrs.

From Figure 10, we can note that when an atom has small content in the sources, it will have fewer remnants (less than 1), that is, it has better fusion effects. By contrast, the atoms owning a large proportion in the sources will have more remnants, that is, their RCRs will overtake their CRAs. This can also be noted in Figure 9 that the red curve gets larger than the blue one when CRA increases.

To sum up, with the aid of the proposed atom-level joint analysis, it was found that the in-focus atoms (with large FMA) usually take large proportions in MFIF sources, and they suffer more serious blur during fusion usually.

To explore the relationship between atom residual ratio (RCR/CRA) and fusion quality, we compute the average residual ratio of the top 50 atoms which have the largest CRA in the 6th group of source images. The analysis result indicates that the algorithms owning better objective evaluation also have less atom residual ratios.

6 | DISCUSSION

In the above experiments, five contrasted metrics can be divided into two categories: Information theory-based MI and NMI, similarity comparison-based $Q^{IB/F}$, VIFF and the proposed metric. The first two are based on mutual entropy, while the later three differentiate from each other by the local features they used: $Q^{IB/F}$ employs simple edge contents, VIFF mainly depends on spatially local SNR, and our metric rely on atom absorption degree. Further discussions are given below:

1. MI, NMI, $Q^{IB/F}$ and our metric show a high degree of consistency although they have different principles. The experiments in Section 5.2.3 demonstrate the correspondence between the objective measurement and subjective observation.
2. Compared with other metrics, not only can our method provide a holistic performance assessment but also can provide deep analysis on fusion effect. This is highly desirable for the design and improvement of SR-based fusion algorithms. Our experiments demonstrate that for the same dictionary, some atoms always own a larger proportion in different sources, so these atoms have more remnants after fusion accordingly. From the point of view of atom fusion, the aim of MFIF algorithms is to decrease the residual degree of in-focus atoms while retaining low total residual degree. So, the method to decrease the RCRs of in-focus atoms should be more emphasised in MFIF.
3. Among our tests, the evaluation results of VIFF are significantly different from the others. NSCT-SR that combines SR and MST gets the best VIFF evaluation almost in all experiments. The reason behind this may be that they all depend on multi-scale analysis, so VIFF can exhibit the advantages of this combinational method better.
4. A global learning dictionary was employed for the convenience of comparisons. An adaptive-learning dictionary trained with the images to be fused also presents the same effectiveness in experiments.
5. Our tests demonstrate that when only a part of patch vectors equidistantly picked were used for quality evaluation, the accuracy of evaluation results are also held. For the 256*256 sources, our experiments indicate that when only 1/8 patches are employed to carry out the measurement, the variation amplitudes of our metric are less than 3% compared with the measurements using all patches. This can effectively decrease the calculation cost of our method.

7 | CONCLUSION

In this study, inspired by SR model and JSR concept, we develop an atom similarity comparison-based evaluation method for MFIF. Our main work of the study includes:

1. An SR-based metric was designed to assess the performance of MFIF algorithms. JSR model is used to extract atom remnants of the source images. The atom remnants and the
corresponding atom focus measure are combined to assess MFIF performance.

2. The correctness of the proposed metric was contrasted with other four commonly used metrics through several fusion experiments. The objective and subjective comparisons demonstrated the correctness of our method.

3. The detailed analyses of the characteristics of atom remnants are reported. The results of our study indicate that the residual degree of atoms is related to three factors: Atom contents in the source images, fusion algorithms used and atom focus measure. From the point of view of SR, MFIF algorithms need to focus on reducing the residual degree of in-focus atoms for achieving better fusion quality.

ACKNOWLEDGEMENTS
The authors would like to thank the anonymous reviewers for their valuable comments. This work was supported by the National Natural Science Foundation of China under Grant 61274021.

REFERENCES
1. Li, X., et al.: Multi-focus image fusion algorithm based on multilevel morphological component analysis and support vector machine. IET Image Proc. 11(10), 919–926 (2017)
2. Li, S., et al.: Pixel-level image fusion: A survey of the state of the art. Inf. Fusion 33, 100–112 (2017)
3. Zhang, Q., et al.: Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review. Inf. Fusion 40, 57–75 (2018)
4. Liu, Z., et al.: Objective assessment of multi-resolution image fusion algorithms for context enhancement in night vision: A comparative study. IEEE Trans. Pattern Anal. Mach. Intell. 34(1), 94–109 (2012)
5. Li, S., et al.: Performance comparison of different multi-resolution transforms for image fusion. Inf. Fusion 12(2), 74–84 (2011)
6. Zha, Z., et al.: Compressed sensing image reconstruction via adaptive sparse nonlocal regularization. Visual Comput. 34(1), 117–137 (2018)
7. Cheng, H.: Sparse representation: Modeling and learning in visual recognition–theory, algorithms and applications. Springer, London (2015)
8. Duarte, M., et al.: Distributed compressed sensing of jointly sparse signals. In: Conference Record of the Thirty-Ninth Asilomar Conference on Signals Systems and Computers, Pacific Grove, California, pp. 1537–1541 (2005)
9. Yang, Y., et al.: Multi-focus image fusion based on a non-fixed-base dictionary and multi-measure optimization. IEEE Access 7(4), 46376–46388 (2019)
10. Xiang, F., et al.: Robust image fusion with block sparse representation and online dictionary learning. IET Image Proc. 12(3), 345–353 (2018)
11. Zhang, Q., et al.: Robust multi-focus image fusion using multi-task sparse representation and spatial context. IEEE Trans. Image Process. 25(5), 2045–2058 (2016)
12. He, G., et al.: Image fusion method based on simultaneous sparse representation with non-subsampled contourlet transforms. IET Comput. Vision 13(2), 240–248 (2019)
13. Liu, Y., et al.: A general framework for image fusion based on multi-scale transform and sparse representation. Inf. Fusion 24, 147–164 (2015)
14. Nejati, M., et al.: Multi-focus image fusion using dictionary-based sparse representation. Inf. Fusion 25, 72–84 (2014)
15. Li, H., et al.: Joint medical image fusion, denoising and enhancement via discriminative low-rank sparse dictionaries learning. Pattern Recognit. 79, 130–146 (2018)
16. Liu, Y., et al.: Deep learning for pixel-level image fusion: Recent advances and future prospects. Inf. Fusion 42, 158–173 (2018)
17. Qu, G., et al.: Information measure for performance of image fusion. Electron. Lett. 38(7), 313–315 (2002)
18. Hossny, M., et al.: Comments on information measure for performance of image fusion. Electron. Lett. 44(18), 1066–1067 (2008)
19. Xydeas, C., Petrovic, V.: Objective image fusion performance measure. Electron. Lett. 36(4), 308–309 (2000)
20. Han, Y., et al.: A new image fusion performance metric based on visual information fidelity. Inf. Fusion 14(2), 127–135 (2013)
21. Wang, Z., et al.: Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. 13(4), 600–612 (2004)
22. Hu, Y., Zhang, B., A bio-inspired two-scale complementarity evaluation method for image fusion. Int. J. Wavelets Multiresolution Inf. Process. 15(6), 1750053 (2017)
23. Piella, G., Heijmans, H.: A new quality metric for image fusion. In: IEEE International Conference on Image Processing, Barcelona, Spain, pp. 173–176 (2003)
24. Huang, W., Jing, Z.: Evaluation of focus measures in multi-focus image fusion. Pattern Recognit. Lett. 28(4), 493–500 (2007)
25. Kumar, B.: Image fusion based on pixel significance using cross bilateral filter. Signal Image Video Process. 9(5), 1193–1204 (2015)
26. Liu, Y., et al.: Multi-focus image fusion with a deep convolutional neural network. Inf. Fusion 36, 191–207 (2017)
27. Liu, Y., et al.: Image fusion with convolutional sparse representation. IEEE Signal Process. Lett. 23(12), 1882–1886 (2016)
28. Kumar, B.: Multi-focus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform. Signal Image Video Process. 7(6), 1125–1143 (2013)
29. Liu, Y., et al.: Multi-focus image fusion with dense SIFT. Inf. Fusion 23, 139–155 (2015)
30. Li, S., et al.: Image fusion with guided filtering. IEEE Trans. Image Process. 22(7), 2864–2875 (2013)
31. Li, S., et al.: Image matting for fusion of multi-focus images in dynamic scenes. Inf. Fusion 14(2), 147–162 (2013)

How to cite this article: Hu Y, Zhang B, Zhang J, Gao Q. Multi-focus image fusion evaluation based on jointly sparse representation and atom focus measure. IET Image Process. 2021;15:1032–1041. https://doi.org/10.1049/ipr2.12083