Information Fusion 24 (2015) 147–164

Contents lists available at ScienceDirect

Information Fusion

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

A general framework for image fusion based on multi-scale transform and sparse representation

Yu Liu [a], Shuping Liu [a], Zengfu Wang [a,b,*]

[a] Department of Automation, University of Science and Technology of China, Hefei 230026, China
[b] Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China

ARTICLE INFO

Article history:
Received 28 January 2014
Received in revised form 1 September 2014
Accepted 9 September 2014
Available online 5 October 2014

Keywords:
Image fusion
Multi-scale transform
Sparse representation

ABSTRACT

In image fusion literature, multi-scale transform (MST) and sparse representation (SR) are two most widely used signal/image representation theories. This paper presents a general image fusion framework by combining MST and SR to simultaneously overcome the inherent defects of both the MST- and SR-based fusion methods. In our fusion framework, the MST is firstly performed on each of the pre-registered source images to obtain their low-pass and high-pass coefficients. Then, the low-pass bands are merged with a SR-based fusion approach while the high-pass bands are fused using the absolute values of coefficients as activity level measurement. The fused image is finally obtained by performing the inverse MST on the merged coefficients. The advantages of the proposed fusion framework over individual MST- or SR-based method are first exhibited in detail from a theoretical point of view, and then experimentally verified with multi-focus, visible-infrared and medical image fusion. In particular, six popular multi-scale transforms, which are Laplacian pyramid (LP), ratio of low-pass pyramid (RP), discrete wavelet transform (DWT), dual-tree complex wavelet transform (DTCWT), curvelet transform (CVT) and nonsubsampled contourlet transform (NSCT), with different decomposition levels ranging from one to four are tested in our experiments. By comparing the fused results subjectively and objectively, we give the best-performed fusion method under the proposed framework for each category of image fusion. The effect of the sliding window's step length is also investigated. Furthermore, experimental results demonstrate that the proposed fusion framework can obtain state-of-the-art performance, especially for the fusion of multimodal images.

1. Introduction

In recent years, image fusion has become an important issue in image processing community. The target of image fusion is to generate a composite image by integrating the complementary information from multiple source images of the same scene [1]. For an image fusion system, the input source images can be acquired from either different types of imaging sensors or a sensor whose optical parameters can be changed, and the output called fused image will be more suitable for human or machine perception than any individual source image. Image fusion technique has been widely employed in many applications such as computer vision, surveillance, medical imaging, and remote sensing.

Multi-scale transform (MST) theories are the most popular tools used in various image fusion scenarios such as multi-focus image fusion, visible-infrared image fusion, and multimodal medical image fusion. Classical MST-based fusion methods include pyramid-based ones like Laplacian pyramid (LP) [2], ratio of low-pass pyramid (RP) [3] and gradient pyramid (GP) [4], wavelet-based ones like discrete wavelet transform (DWT) [5], stationary wavelet transform (SWT) [6] and dual-tree complex wavelet transform (DTCWT) [7], and multi-scale geometric analysis (MGA)-based ones like curvelet transform (CVT) [8] and nonsubsampled contourlet transform (NSCT) [9]. In general, the MST-based fusion methods consist of the following three steps [10]. First, decompose the source images into a multi-scale transform domain. Then, merge the transformed coefficients with a given fusion rule. Finally, reconstruct the fused image by performing the corresponding inverse transform over the merged coefficients. These methods assume that the underlying salient information of the source images can be extracted from the decomposed coefficients. Obviously, the selection of transform domain plays a crucial role

* Corresponding author at: Department of Automation, University of Science and Technology of China, Hefei 230026, China. Tel.: +86 551 63600634.
E-mail addresses: liuyu1@mail.ustc.edu.cn (Y. Liu), fengya@mail.ustc.edu.cn (S. Liu), zfwang@ustc.edu.cn (Z. Wang).

http://dx.doi.org/10.1016/j.inffus.2014.09.004
1566-2535/

in these methods. A comparative study of different MST-based methods is reported in [11], where Li et al. found that the NSCT-based method can generally achieve the best results. In addition to the selection of transform domain, the fusion rule in either high-pass or low-pass band also has a great impact on the fused results. Conventionally, the absolute value of high-pass coefficient is used as the activity level measurement for high-pass fusion. The simplest rule is selecting the coefficient with largest absolute value at each pixel position (the "max-absolute" rule). Many improved high-pass fusion rules which make use of the neighbor coefficients' information have also been developed. However, compared with the great concentration on developing effective rules for high-pass fusion, less attention has been paid to the fusion of low-pass bands. In most MST-based fusion methods, the low-pass bands are just simply merged by averaging all the source inputs (the "averaging" rule).

Sparse representation addresses the signals' natural sparsity, which is in accord with the physiological characteristics of human visual system [12]. The basic assumption behind SR is that a signal $\mathbf{x} \in \mathbf{R}^n$ can be approximately represented by a linear combination of a "few" atoms from an overcomplete dictionary $\mathbf{D} \in \mathbf{R}^{n\times m}(n < m)$, where n is the signal dimension and m is the dictionary size. That is, the signal $\mathbf{x}$ can be expressed as $\mathbf{x} \approx \mathbf{D}\alpha$, where $\alpha \in \mathbf{R}^m$ is the unknown sparse coefficient vector. As the dictionary is overcomplete, there are numerous feasible solutions for this underdetermined system. The target of SR is to calculate the sparsest α which contains the fewest nonzero entries among all feasible solutions (known as sparse coding). In SR-based image processing methods, the sparse coding technique is often performed on local image patches for the sake of algorithm stability and efficiency [13]. Yang and Li [14] first introduced SR into image fusion. The sliding window technique (patches are overlapped) is adopted in their method to make the fusion process more robust to noise and misregistration. In [14], the sparse coefficient vector is used as the activity level measurement. Particularly, among all the source sparse vectors, the one owning the maximal $l_1 - norm$ is selected as the fused sparse vector (the "max-L1" rule). The fused image is finally reconstructed with all the fused sparse vectors. Their experimental results show that the SR-based fusion method owns clear advantages over traditional MST-based methods for multi-focus image fusion, and can lead to state-of-the-art results. In the past few years, the SR-based fusion has emerged as a new active branch in image fusion research with many improved approaches being proposed [15–18].

Although both the MST- and SR-based methods have achieved great success in image fusion, it is worthwhile to notice that both of them have some defects, which will be further discussed in this paper. Moreover, to overcome the related disadvantages, we present a general image fusion framework by taking the complementary advantages of MST and SR. Specifically, the low-pass MST bands are merged with a SR-based fusion approach while the high-pass MST bands are fused using the conventional "max-absolute" rule with a local window based consistency verification scheme [5]. To verify the effectiveness of the proposed framework, six popular multi-scale transforms (MSTs), which are LP, RP, DWT, DTCWT, CVT and NSCT, with different decomposition levels ranging from one to four are tested in our experiments. By comparing the fused results subjectively and objectively, we give the best-performed methods under the proposed framework for the fusion of multi-focus, visible-infrared and medical images, respectively. The effect of the sliding window's step length is also investigated. Experimental results demonstrate that the combined methods can clearly outperform both the MST- and SR-based methods. Furthermore, the proposed fusion methods can obtain state-of-the-art fused results, especially for the fusion of medical images as well as visible-infrared images.

The rest of this paper is organized as follows. We first present the detailed fusion framework in Section 2. In Section 3, the disadvantages of MST- and SR-based methods and why the proposed framework can overcome them are discussed from a theoretical perspective. The experimental results are given in Section 4. Section 5 summarizes some main conclusions of this paper.

2. Proposed fusion framework

To better exhibit the advantages of the proposed framework over MST- and SR-based methods, we first present the details of our framework in this section.

2.1. Dictionary learning

The overcomplete dictionary determines the signal representation ability of sparse coding. Generally, there are two main categories of offline approaches to obtain a dictionary. The first one is directly using the analytical models such as discrete cosine transform (DCT) and CVT. However, this category of dictionary is restricted to signals of a certain type and cannot be used for an arbitrary family of signals. The second category is applying the machine learning technique to obtain the dictionary from a large number of training image patches. Suppose that M training patches of size $\sqrt{n} \times \sqrt{n}$ are rearranged to column vectors in the $\mathbf{R}^n$ space, thereby the training database $\{\mathbf{y}_i\}_{i=1}^M$ is constructed with each $\mathbf{y}_i \in \mathbf{R}^n$. The dictionary learning model can be presented as

$$\min_{\mathbf{D},\{\alpha_i\}_{i=1}^M} \sum_{i=1}^{M} \|\alpha_i\|_0 \quad s.t. \quad \|\mathbf{y}_i - \mathbf{D}\alpha_i\|_2 < \varepsilon, \quad i \in \{1,\ldots,M\}, \tag{1}$$

where $\varepsilon > 0$ is an error tolerance, $\{\alpha_i\}_{i=1}^M$ is the unknown sparse vectors corresponding to $\{\mathbf{y}_i\}_{i=1}^M$ and $\mathbf{D} \in \mathbf{R}^{n\times m}$ is the unknown dictionary to be learned. Some effective methods such as MOD [19] and K-SVD [20] have been proposed to solve this problem. The learned dictionaries usually have better representation ability than the pre-constructed ones, so we adopt the learning-based approach in this paper.

In this work, the sparse coding technique is employed for the fusion of MST low-pass bands. One possible way to get the training patches is sampling from the corresponding MST low-pass bands which are obtained from some training images under the same decomposition condition. However, in this case, the dictionary learning process should be repeated once either the selected transform domain or even one specific parameter (such as the decomposition level or selected image filter) is changed. Obviously, this will decrease the flexibility and practicality of the fusion method to a large extent. In this paper, we aim to learn a universal dictionary which can be used in any specific transform domain and parameter settings. As is well known, the MST low-pass band obtained by image filtering can be viewed as a smooth version of the original image. Since the numerous "flat" patches contained in a natural image can be well sparsely represented by a dictionary learned from natural image patches, it is theoretically feasible to use the same dictionary to represent the patches in the low-pass bands so long as the mean value of each sampled patch is subtracted to zero before training. In this situation, the mean value of each atom in the obtained dictionary is also zero, so the atoms only contain structural information. For an input patch to be represented, its mean value should also be subtracted to zero before sparse coding. Thus, we can directly use natural image patches to learn a universal dictionary.