Information Fusion 36 (2017) 191–207

Contents lists available at ScienceDirect

Information Fusion

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

Multi-focus image fusion with a deep convolutional neural network

Yu Liu[a], Xun Chen[a,*], Hu Peng[a], Zengfu Wang[b]

[a] Department of Biomedical Engineering, Hefei University of Technology, Hefei 230009, China
[b] Department of Automation, University of Science and Technology of China, Hefei 230026, China

ARTICLE INFO

Article history:
Received 21 March 2016
Revised 29 November 2016
Accepted 4 December 2016
Available online 5 December 2016

Keywords:
Image fusion
Multi-focus image fusion
Deep learning
Convolutional neural networks
Activity level measurement
Fusion rule

ABSTRACT

As is well known, activity level measurement and fusion rule are two crucial factors in image fusion. For most existing fusion methods, either in spatial domain or in a transform domain like wavelet, the activity level measurement is essentially implemented by designing local filters to extract high-frequency details, and the calculated clarity information of different source images are then compared using some elaborately designed rules to obtain a clarity/focus map. Consequently, the focus map contains the integrated clarity information, which is of great significance to various image fusion issues, such as multi-focus image fusion, multi-modal image fusion, etc. However, in order to achieve a satisfactory fusion performance, these two tasks are usually difficult to finish. In this study, we address this problem with a deep learning approach, aiming to learn a direct mapping between source images and focus map. To this end, a deep convolutional neural network (CNN) trained by high-quality image patches and their blurred versions is adopted to encode the mapping. The main novelty of this idea is that the activity level measurement and fusion rule can be jointly generated through learning a CNN model, which overcomes the difficulty faced by the existing fusion methods. Based on the above idea, a new multi-focus image fusion method is primarily proposed in this paper. Experimental results demonstrate that the proposed method can obtain state-of-the-art fusion performance in terms of both visual quality and objective assessment. The computational speed of the proposed method using parallel computing is fast enough for practical usage. The potential of the learned CNN model for some other-type image fusion issues is also briefly exhibited in the experiments.

1. Introduction

In the field of digital photography, it is often difficult for an imaging device like a digital single-lens reflex camera to take an image in which all the objects are captured in focus. Typically, under a certain focal setting of optical lens, only the objects within the depth-of-field (DOF) have sharp appearance in the photograph while other objects are likely to be blurred. A popular technique to obtain an all-in-focus image is fusing multiple images of the same scene taken with different focal settings, which is known as multi-focus image fusion. At the same time, multi-focus image fusion is also an important subfield of image fusion. With or without modification, many algorithms for merging multi-focus images can also be employed for other image fusion tasks such as visible-infrared image fusion and multi-modal medical image fusion (and vice versa). From this point of view, the meaning of studying multi-focus image fusion is twofold, which makes it an active topic in image processing community. In recent years, various image fusion methods have been proposed, and these methods can be roughly classified into two categories [1]: transform domain methods and spatial domain methods.

The most classic transform domain fusion methods are based on multi-scale transform (MST) theories, which have been applied in image fusion for more than thirty years since the Laplacian pyramid (LP)-based fusion method [2] was proposed. Since then, a large number of multi-scale transform based image fusion methods have appeared in this field. Some representative examples include the morphological pyramid (MP)-based method [3], the discrete wavelet transform (DWT)-base method [4], the dual-tree complex wavelet transform (DTCWT)-based method [5], and the non-subsampled contourlet transform (NSCT)-based method [6]. These MST-based methods share a universal three-step framework, namely, decomposition, fusion and reconstruction [7]. The basic assumption of MST-based methods is that the activity level of source images can be measured by the decomposed coefficients in a selected transform domain. Apart from the selection of MST domain,

* Corresponding author.
E-mail addresses: yuliu@hfut.edu.cn, liuyu1@mail.ustc.edu.cn (Y. Liu), xun.chen@hfut.edu.cn (X. Chen).

http://dx.doi.org/10.1016/j.inffus.2016.12.001

the rules designed for merging decomposed coefficients also play a very important role in MST-based methods, and many studies have also been taken in this direction [8–11]. In recent years, a new kind of transform domain fusion methods [12–16] has emerged as an attractive branch in this field. Different from the above introduced MST-based methods, these methods transform images into a single-scale feature domain with some advanced signal representation theories such as independent component analysis (ICA) and sparse representation (SR). This category of methods usually employs the sliding window technique to pursue an approximate shift-invariant fusion process. The key issue of these methods is to explore an effective feature domain for the calculation of activity level. For instance, as one of the most representative approaches belonging to this category, the SR-based method [13] transforms the source image patches into sparse domain and applies the L1-norm of sparse coefficients as the activity level measurement.

The spatial domain methods in the early stage usually adopt a block-based fusion strategy, in which the source images are decomposed into blocks and each pair of block is fused with a designed activity level measurement like spatial frequency and sum-modified-Laplacian [17]. Clearly, the block size has a great impact on the quality of fusion results. Since the earliest block-based methods [18,19] using manually fixed size appeared, many improved versions have been proposed on this topic, such as the adaptive block based method [20] using differential evolution algorithm to obtain a fixed optimal block size, and some recently introduced quad-tree based methods [21,22] in which the images can be adaptively divided into blocks with different sizes according to image content. Another type of spatial domain methods [23,24] is based on image segmentation by sharing the similar idea of block-based methods, but the fusion quality of these methods relies heavily on the segmentation accuracy. In the past few years, some novel pixel-based spatial domain methods [25–31] based on gradient information have been proposed, which can currently obtain state-of-the-art results in multi-focus image fusion. To further improve the fusion quality, these methods usually apply relatively complex fusion schemes (can be regarded as *rules* in a broad sense) to their calculation results of activity level measurement.

It is well known that for either transform domain or spatial domain image fusion methods, activity level measurement and fusion rule are two crucial factors. In most existing image fusion methods, these two issues are considered separately and designed manually [32]. To make further improvements, many recently proposed methods tend to be more and more complicated on these two issues. In the MST-based methods, new transform domains in [33,34] and new fusion rules in [9–11] were introduced. In the SR-based methods, there were new sparse models and more complex fusion rules in [35–37]. In the block-based methods, new focus measures were proposed in [21,22]. In the pixel-based methods, new activity level measurements were introduced in [27,29] and the fusion schemes employed in [26,28–30] are very intricate. The above introduced works were all published within the last five years. It is worthwhile to clarify that we don't mean these elaborately designed activity level measurements and fusion rules are not important contributions, but the problem is that manual design is really not an easy task. Moreover, from a certain point of view, it is almost impossible to come up with an ideal design that takes all the necessary factors into account.

In this paper, we address this problem with a deep learning approach, aiming to learn a direct mapping between source images and focus map. The focus map here indicates a pixel-level map which contains the clarity information after comparing the activity level measure of source images. To achieve this target, a deep convolutional neural network (CNN) [38] trained by high-quality image patches and their blurred versions is adopted to encode the mapping. The main novelty of this idea is that the activity level measurement and fusion rule can be jointly generated through learning a CNN model, which overcomes the above difficulty faced by existing fusion methods. Based on this idea, we propose a new multi-focus image fusion method in spatial domain. We demonstrate that the focus map obtained from the convolutional network is reliable that very simple consistency verification techniques can lead to high-quality fusion results. The computational speed of the proposed method using parallel computing is fast enough for practical usage. At last, we briefly exhibit the potential of the learned CNN model for some other-type image fusion issues, such as visible-infrared image fusion, medical image fusion and multi-exposure image fusion.

To the best of our knowledge, this is the first time that the convolutional neural network is applied to an image fusion task. The most similar work was proposed by Li et al. [19], in which they pointed out that the multi-focus image fusion can be viewed as a classification problem and presented a fusion method based on artificial neural networks. However, there exist significant differences between the method in [19] and our method. The method in [19] first calculates three commonly used focus measures (feature extraction) and then feeds them to a three-layer (input-hidden-output) network, so the network just acts as a classifier for the fusion rule design. As a result, the source images must be fused patch by patch in [19]. In this work, the CNN model is simultaneously used for activity level measure (feature extraction) and fusion rule design (classification). The original image content are the input of the CNN model. Thus, the network in this study should be deeper than the "shallow" network used in [19]. Considering that the GPU parallel computation is becoming more and more popular, the computational speed of CNN-based fusion is not a concern nowadays. In addition, owing to the convolutional characteristic of CNNs [39], the source images in our method can be fed to the network as a whole to further improve the computational efficiency.

The rest of this paper is organized as follows. In Section 2, we give a brief introduction to CNN and explain its feasibility as well as advantage for image fusion problem. In Section 3, the proposed CNN-based multi-focus fusion method is presented in detail. The experimental results and discussions are provided in Section 4. Finally, Section 5 concludes the paper.

2. CNN model for image fusion

2.1. CNN model

CNN is a typical deep learning model, which attempts to learn a hierarchical feature representation mechanism for signal/image data with different levels of abstraction [40]. More concretely, CNN is a trainable multi-stage feed-forward artificial neural network and each stage contains a certain number of *feature maps* corresponding to a level of abstraction for features. Each unit or coefficient in a feature map is called a *neuron*. The operations such as linear convolution, non-linear activation and spatial pooling applied to neurons are used to connect the feature maps at different stages.

Local receptive fields, shared weights and *sub-sampling* are three basic architectural ideas of CNNs [38]. The first one indicates a neuron at a certain stage is only connected with a few spatially neighboring neurons at its previous stage, which is in accord with the mechanism of mammal visual cortex. As a result, local convolutional operation is performed on the input neurons in CNNs, unlike the fully-connected mechanism used in conventional multilayer perception. The second idea means the weights of a convolutional kernel is spatially invariant in feature maps at a certain stage. By combining these two ideas, the number of weights to be trained is greatly reduced. Mathematically, let x^i and y^j denote the i-th input feature map and j-th output feature map of a convolu-