Neurocomputing 202 (2016) 12–19

Contents lists available at ScienceDirect

Neurocomputing

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

Infrared and visible image fusion using total variation model

Yong Ma [a], Jun Chen [b], Chen Chen [c], Fan Fan [a,*], Jiayi Ma [a]

[a] Electronic Information School, Wuhan University, Wuhan 430072, China
[b] School of Automation, China University of Geosciences, Wuhan 430074, China
[c] Department of Electrical and Computer Engineering, UIUC, Urbana, IL 61801, USA

ARTICLE INFO

Article history:
Received 25 November 2015
Received in revised form
2 March 2016
Accepted 8 March 2016
Available online 28 March 2016

Keywords:
Image fusion
Infrared
Total variation

ABSTRACT

Image fusion is a process of combining complementary information from multiple images of the same scene into an image, so that the resultant image contains a more accurate description of the scene than any of the individual source images. In this paper, we propose a novel fusion strategy for infrared (IR) and visible images based on total variation (TV) minimization. By constraining the fused image to have similar pixel intensities with the IR image and similar gradients with the visible image, it tends to simultaneously keep the thermal radiation and appearance information in the source images. We evaluate our method on a publicly available database with comparisons to other seven fusion methods. Our results have a major difference that the fused images look like sharpened IR images with detailed appearance information. The quantitative results demonstrate that our method also can achieve comparable metric values with other state-of-the-art methods.

1. Introduction

Multi-sensor data often provides complementary information about the region surveyed, and image fusion which aims to create new images from such data offering more complex and detailed scene representation has then emerged as a promising research strategy for scene analysis in the areas of remote sensing [1,2], pattern recognition [3,4], medical imaging [5,6] and modern military [7,8]. In this paper, we concentrate on the fusion of multi-sensor data such as thermal infrared (IR) and visible images, which can lead to better performance for human visual perception, object detection, as well as target recognition [9,10].

The goal of image fusion is to identify the most important information in the source images and to transfer, without distortion or loss, this information into a fused image. For the problem of IR and visible image fusion, visible sensors capture reflected lights with abundant appearance information, and it is better for establishing a discriminative model. In contrast, IR sensors capture principally thermal radiations emitted by objects, which are not affected by illumination variation or disguise and hence, it can overcome some of the obstacles to discover the target and work day and night. However, IR image has lower spatial resolution than visible image, where appearance features such as textures in a visible image often get lost in the corresponding IR image since textures seldom influence the heat emitted by an object. Therefore, it is beneficial for automatic target detection and unambiguous localization to fuse the thermal radiation and texture information into a single image.

The process of image fusion can be performed at different levels depending on the information representation and applications. A common categorization is to distinguish between pixel, feature and symbol levels [11]. Fusion at pixel-level represents fusion at the lowest level referring to combining the raw source images into a single image [12]. Fusion at higher level such as feature-level or symbol level combines information in the form of feature descriptors and probabilistic variables [13]. However, pixel-level fusion is still a popular strategy for most image fusion applications, as it has the main advantage that the original measured quantities are directly involved in the fusion process. Besides, pixel-level fusion algorithms are computationally efficient and easy to implement [14]. In this paper, we only focus on the pixel-level image fusion problem.

A prerequisite for pixel-level fusion is that multi-sensor images have to be correctly registered on a pixel-by-pixel basis. Image registration techniques have been discussed extensively in the literature [15–19]. Throughout this paper, it will be assumed that all source images have been registered. To address the pixel-level fusion problem, many methods have been proposed in the past decades [20–31]. The simplest strategy is to take the average of the

* Corresponding author.
E-mail addresses: mayong@whu.edu.cn (Y. Ma), junchen@cug.edu.cn (J. Chen), cchen156@illinois.edu (C. Chen), fanfan@whu.edu.cn (F. Fan), jyma2010@gmail.com (J. Ma).

http://dx.doi.org/10.1016/j.neucom.2016.03.009
0925-2312/

source images pixel by pixel. However, such direct method will lead to several undesired side effects including reduced contrast. In order to solve this problem, multi-scale transform (MST) based methods have been proposed which are able to provide much better performance, since they are consistent with human visual system and real-world objects usually consist of structures at different scales [11,32]. Examples of these methods include Laplacian pyramid [20], discrete wavelet transform [33], and nonsubsampled contourlet transform [34]. The MST-based methods have achieved great success in many situations; however, they use the same representations for different source images and try to preserve the same salient features such as edges and lines in the source images. For the problem of IR and visible image fusion, the thermal radiation information in an IR image is characterized by the pixel intensities, and the target typically has larger intensities compared to the background and hence can be easily detected; while the texture information in a visible image is mainly characterized by the gradients, and the gradients with large magnitude (e.g. edges) provide detail information for the scene. Therefore, it is not appropriate to use the same representations for these two types of images during the fusion process. Instead, to preserve the important information as more as possible, the fused image is desirable to keep the main intensity distribution in the IR image and the gradient variation in the visible image. To this end, in this paper we proposed a novel algorithm based on total variation (TV) minimization for IR and visible image fusion.

More precisely, we formulate the fusion as a TV minimization problem, where the data fidelity term constrains that the fused image should have the similar pixel intensities with the given IR image, and the regularization term ensures that the gradient distribution in the visible image can be transferred into the fused image. The ℓ^1 norm is employed to encourage the sparseness of the gradients [35], and the optimization problem can then be solved via existing TV minimization techniques [36].

The contribution of this paper is two-fold. On the one hand, we propose a new IR and visible image fusion algorithm based on total variation minimization. It tends to simultaneously preserve the thermal radiation information as well as the detailed appearance information in the source images, and to the best of our knowledge, such fusion strategy has not yet been studied. On the other hand, we provide both qualitative and quantitative comparisons with several state-of-the-art approaches on a publicly available dataset. Compared to previous methods, our method can generate fusion results looking like sharpened IR images with detailed scene representation and hence, it is able to improve the reliability of automatic target detection and recognition systems.

The rest of the paper is organized as follows. Section 2 presents the formulation of the proposed fusion algorithm for IR and visible image fusion. In Section 3, we demonstrate our method for fusion on a publicly available dataset with comparisons to several state-of-the-art approaches, followed by some concluding remarks in Section 4.

2. Method

In this section, we present the layout of our IR and visible image fusion method. To this end, we first introduce the TV minimization problem, and then present our fusion method based on TV minimization.

2.1. Total variation minimization

The TV model was proposed by Beck et al. [37] as a regularizing criterion to solve the image denoising problem due to its property of effectively preserving edge information, which has evolved from an image denoising method [38] into a more general technique for inverse problems [39], including deblurring [40], blind deconvolution [41], inpainting [42], super-resolution [43], texture analysis [44] and smoothing [45]. For an image of size $m \times n$, we denote by $\mathbf{u} \in \mathbb{R}^{mn\times 1}$ the column-vector form of its pixel intensities, which has gray-scale values ranging from 0 to 255. The TV model of $\mathbf{u}$ is defined as follows:

$$J(\mathbf{u}) = \sum_{i=1}^{mn} |\nabla_i \mathbf{u}| = \sum_{i=1}^{mn} \sqrt{(\nabla_i^h \mathbf{u})^2 + (\nabla_i^v \mathbf{u})^2}, \tag{1}$$

where $|u| := \sqrt{u_1^2 + u_2^2}$ for every $u = (u_1, u_2) \in \mathbb{R}^2$, $\nabla_i = (\nabla_i^h, \nabla_i^v)$ denotes the image gradient ∇at pixel i with ∇^{h} and ∇^{v} being linear operators corresponding to the horizontal and vertical first-order differences, respectively. More specifically, $\nabla_i^h \mathbf{u} = \mathbf{u}_i - \mathbf{u}_{r(i)}$ and $\nabla_i^v \mathbf{u} = \mathbf{u}_i - \mathbf{u}_{b(i)}$, where $r(i)$ and $b(i)$ represent the nearest neighbor to the right and below the pixel i. Besides, if pixel i is located in the last row or column, $r(i)$ and $b(i)$ are both set to be i.

With the TV model in Eq. (1), many image processing tasks can be formulated as the following inverse problem with regularization constraint:

$$\mathbf{u}^* := \arg\min_{\mathbf{u}} \frac{1}{2} \| \mathbf{u} - \mathbf{f} \|^2 + \lambda \mathbf{J}(\mathbf{u}), \tag{2}$$

where the first term $\| \mathbf{u} - \mathbf{f} \|^2$ is the data fidelity item, which stands for the fidelity between the observed image $\mathbf{f}$ and the original unknown image $\mathbf{u}$. The total variation $J(\mathbf{u})$ in the second term plays a role of regularization. λ is the regularization parameter controlling the tradeoff between the data fidelity and regularization item. Eq. (2) is the classic TV minimization problem which can be efficiently solved by using existing algorithms [36]. And it has been investigated for solving many image processing tasks such as denoising, deblurring, and reconstruction.

2.2. The proposed fusion method

Given a pair of aligned IR and visible images, our goal is to generate a fused image that simultaneously preserves the thermal radiation information and the detailed appearance information in the two images, respectively. Here the IR, visible and fused images are all supposed to be gray scale images of size $m \times n$, and their column-vector forms are respectively denoted by $\mathbf{u}$, $\mathbf{v}$, $\mathbf{x} \in \mathbb{R}^{mn\times 1}$ with gray-scale values ranging from 0 to 255. Typically, the thermal radiation is characterized by the pixel intensities, and then the fused image is expected to have the similar pixel intensities with the IR image, for example, the following empirical error should be as small as possible

$$\mathcal{E}_1(\mathbf{x}) = \tfrac{1}{2} \|\mathbf{x} - \mathbf{u}\|_2^2. \tag{3}$$

To fuse the detailed appearance information, a straightforward scenario is to require the fused image also to have the similar pixel intensities with the visible image. However, the intensity of a pixel in the same physical location may be significantly different for IR and visible images, as they are manifestations of two different phenomena and hence, it is not appropriate to generate $\mathbf{x}$ by simultaneously minimizing $\|\mathbf{x} - \mathbf{u}\|_2^2$ and $\|\mathbf{x} - \mathbf{v}\|_2^2$. Note that the detailed appearance information about the scene is essentially characterized by the gradients in the image. Therefore, we propose to constrain the fused image to have similar pixel gradients rather than similar pixel intensities with the visible image. As visible images are often piece-wise smooth, their gradients tend to be sparse and gradients with large magnitude correspond to the edges. It is widely known that the ℓ^1 norm encourages sparsity and ℓ^2 norm does not, thus we consider minimizing the gradient differences with ℓ^1 norm to encourage sparseness of the gradients:

$$\mathcal{E}_2(\mathbf{x}) = \| \nabla \mathbf{x} - \nabla \mathbf{v} \|_1. \tag{4}$$