Information Fusion 31 (2016) 100–109

Contents lists available at ScienceDirect

Information Fusion

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

Infrared and visible image fusion via gradient transfer and total variation minimization

Jiayi Ma[a], Chen Chen[b], Chang Li[c], Jun Huang[a,*]

[a] Electronic Information School, Wuhan University, Wuhan 430072, China
[b] Department of Electrical and Computer Engineering, UIUC, Urbana, IL 61801, United States
[c] School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China

ARTICLE INFO

Article history:
Received 8 June 2015
Revised 24 October 2015
Accepted 2 February 2016
Available online 9 February 2016

Keywords:
Image fusion
Infrared
Registration
Total variation
Gradient transfer

ABSTRACT

In image fusion, the most desirable information is obtained from multiple images of the same scene and merged to generate a composite image. This resulting new image is more appropriate for human visual perception and further image-processing tasks. Existing methods typically use the same representations and extract the similar characteristics for different source images during the fusion process. However, it may not be appropriate for infrared and visible images, as the thermal radiation in infrared images and the appearance in visible images are manifestations of two different phenomena. To keep the thermal radiation and appearance information simultaneously, in this paper we propose a novel fusion algorithm, named *Gradient Transfer Fusion* (GTF), based on gradient transfer and total variation (TV) minimization. We formulate the fusion problem as an ℓ^1-TV minimization problem, where the data fidelity term keeps the main intensity distribution in the infrared image, and the regularization term preserves the gradient variation in the visible image. We also generalize the formulation to fuse image pairs without pre-registration, which greatly enhances its applicability as high-precision registration is very challenging for multi-sensor data. The qualitative and quantitative comparisons with eight state-of-the-art methods on publicly available databases demonstrate the advantages of GTF, where our results look like sharpened infrared images with more appearance details.

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 pattern recognition, remote sensing, medical imaging and modern military [1–3]. Particularly, multi-sensor data such as thermal infrared (IR) and visible images has been used to enhance the performance in terms of human visual perception, object detection, and target recognition [4,5]. For example, 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 less 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 typically 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, which is the main focus of this paper.

Image fusion is a process that can be conducted at varying levels, depending on information representations and applications. These levels are usually categorized into signal, feature, and symbol levels; however, despite their distinctions, they can be combined [6]. Pixel-level fusion, or fusion at the signal level, is the lowest-level type of fusion and involves the combination of raw source images into a single image [7]. By contrast, higher-level fusion involves combining information that usually occurs in the form of feature descriptors and probabilistic variables [8]. For example, source images are usually fused by feature-level algorithms according to the obtained feature properties, such as edges, shapes, textures, or regions. By contrast, symbol-level algorithms fuse a number of symbolic representations based on decision rules that

* Corresponding author. Tel.: +86 13871380842.
E-mail addresses: jyma2010@gmail.com (J. Ma), cchen156@illinois.edu (C. Chen), lichang@hust.edu.cn (C. Li), junhwong@whu.edu.cn (J. Huang).

http://dx.doi.org/10.1016/j.inffus.2016.02.001
1566-2535/

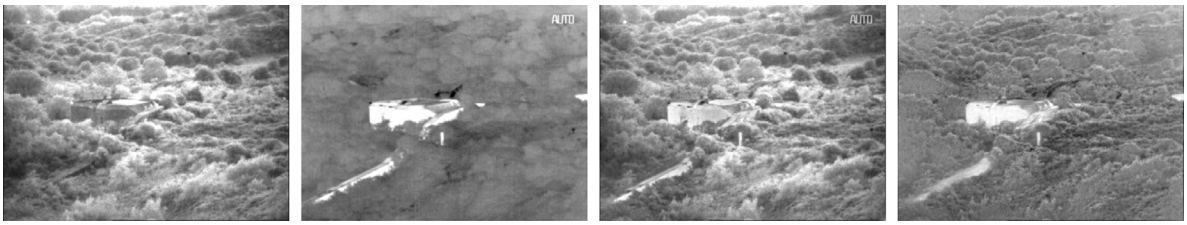

Fig. 1. Schematic illustration of image fusion. From left to right: the visible image, the infrared image, the fusion result of a recent MST-based method LP-SR [18], and the fusion result of our proposed GTF.

promote common interpretation and resolve differences. This study addresses the problem of pixel-level fusion, which is widely used in most image fusion applications, as the original information is contained in the input data and the algorithms are rather easy to implement and time efficient.

In pixel-image fusion, the first problem to be solved is to determine the most important information in the source images to transfer the obtained information into a fused image with the least change possible, especially distortion or loss. To address this problem, many methods have been proposed in the past decades, including methods based on pyramid [9–11], wavelets [12,13], curvelet transform [14], multi-resolution singular value decomposition [15], guided filtering [16], multi-focus [17], sparse representation [18], etc. Averaging the source images pixel by pixel is the simplest strategy. However, a number of undesirable effects, such as reduced contrast, arise from this direct method. In order to solve this problem, multi-scale transform (MST) based methods have been proposed which involve three basic steps: i) the source images are first decomposed into multi-scale representations with low and high frequency information; ii) the multi-scale representations are then fused according to some fusion rules; iii) the inverse transform of composite multi-scale coefficients is finally used to compose the fused image [19]. The MST-based methods are able to provide much better performance as they are consistent with human visual system and real-world objects usually consist of structures at different scales [6]. Examples of these methods include Laplacian pyramid [9], discrete wavelet transform [20], non subsampled contour let transform [21], etc. 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 infrared and visible image fusion, the thermal radiation information in an infrared 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 much as possible, the fused image is desirable to keep the main intensity distribution in the infrared image and the gradient variation in the visible image. To this end, in this paper we proposed a new algorithm, named Gradient Transfer Fusion (GTF), based on gradient transfer and total variation (TV) minimization for infrared and visible image fusion.

More precisely, we formulate the fusion as an optimization problem, where the objective function consists of a data fidelity term and a regularization item. The data fidelity term constrains that the fused image should have the similar pixel intensities with the given infrared image, while 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, and the optimization problem can then be solved via existing ℓ^1-TV minimization techniques [22,23]. To illustrate the main ideas of our method, we show a simple example in Fig. 1. The left two images are the visible and infrared images to be fused, where the visible image contains detailed background and the infrared image highlights the target, i.e. the building. The third image is the fusion result by using a recent MST-based method [18]. We see that the details of the background are kept and the target also becomes brighter. However, it fails to highlight the target as the background is also bright after fusion. This demonstrates the importance of keeping the thermal radiation distribution in the infrared image, and the advantage will be magnified when the scene contains false targets (e.g., decoys) which often occurs in military applications. The rightmost image in Fig. 1 is the fusion result of our GTF algorithm. Clearly, our result preserves the thermal radiation distribution in the infrared image and hence, the target can be easily detected. Meanwhile, the details of the background (i.e. the trees) in the visible image are also kept in our result.

The second problem to be solved in pixel-level image fusion is the accurate alignment of the source images on a pixel-by-pixel basis. This alignment process guarantees that the original information from each source corresponds to the same real-world physical structures. Registration involves various challenges, especially when multi-sensor data, such as IR and visible images, are used. Thus, researchers and scholars have explored the issue of image registration as a process independent from image fusion and have proposed a good number of considerably successful techniques of image registration. In the literature, there are in general two types of approaches for image registration: area-based methods [24,25] and feature-based methods [26–28], as discussed in recent survey papers [29]. The former finds the matching information by using the original pixel intensities in the overlapped region of two images with a specified similarity metric, while the latter seeks correspondence between local features under descriptor similarity and/or spatial geometrical constraints. The area-based methods are preferable in case of few prominent details where the distinctive information is provided by pixel intensities rather than by local shapes and structures, but they suffer from heavy computational complexities, image distortions, as well as illumination changes. By contrast, feature-based methods are more robust to geometric and photometric distortions. As such, the registration problem reduces to determining the correct correspondence and finding the underlying spatial transformation between two sets of extracted features. However, it is difficult to establish accurate alignments for infrared and visible image pairs, since locating reliable features for such images is very challenging and the registration accuracy directly relies on the extracted features. In this paper, we further generalize the proposed GTF so that it is able to fuse unregistered image pairs.

Specifically, we address the registration problem by introducing an additional variable, i.e. the spatial transformation, to the GTF formulation and imposing it on the image with larger field of view. The fused image and the spatial transformation are alternatively optimized until convergence under the assumption that the

other variable is known. To avoid unsatisfying locally optimal solutions and accelerate the convergence, we seek a rough alignment to initialize the transformation by using a feature-based registration method, where the discrete edge map is chosen as the salient features.

Our contribution in this paper includes the following three aspects. (i) We propose a new infrared and visible image fusion algorithm based on gradient transfer and total variation minimization. It can 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. (ii) We generalize the proposed algorithm so that it is able to fuse image pairs without pre-registration, which greatly enhances its applicability as high-precision registration is very challenging for multi-sensor data such as infrared and visible images. (iii) We provide both qualitative and quantitative comparisons with several state-of-the-art approaches on publicly available datasets. Compared to previous methods, our method can generate fusion results looking like sharpened infrared images with highlighted targets and abundant textures 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 and optimization strategy of the proposed GTF algorithm for infrared and visible image fusion. In Section 3, we generalize our GTF to fuse unregistered image pairs, where an iterative method for solving the fused image and spatial transformation is described. Section 4 illustrates our method for fusion on publicly available datasets including both registered and unregistered infrared/visible pairs with comparisons to several state-of-the-art approaches, followed by some concluding remarks in Section 5.

2. The gradient transfer fusion algorithm

In this section, we present the layout of our infrared and visible image fusion method. To this end, we first present our formulation of the fusion problem based on gradient transfer, and then provide the optimization method using total variation minimization.

2.1. Problem formulation

Given a pair of aligned infrared 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 infrared, visible and fused images are all supposed to be gray scale images of size $m \times n$, and their column-vector forms are denoted by $\mathbf{u}, \mathbf{v}, \mathbf{x} \in \mathbb{R}^{mn\times 1}$, respectively.

On the one hand, the thermal radiation is typically characterized by the pixel intensities, and the targets are often distinctly visible in the infrared image, due to the pixel intensity difference between the targets and background. This motivated us to constrain the fused image to have the similar pixel intensity distribution with the given infrared image, for example, the following empirical error measured by some ℓ^p norm ($p \geq 1$) should be as small as possible

$$\mathcal{E}_1(\mathbf{x}) = \frac{1}{p}\|\mathbf{x}-\mathbf{u}\|_p^p. \tag{1}$$

On the other hand, the targets should be depicted in a background from the visual modality to enhance the user's situational awareness. To fuse the detailed appearance information, a straightforward scenario is to require the fused image also has the similar pixel intensities with the visible image. However, the intensity of a pixel in the same physical location may be significantly different for infrared and visible images, as they are manifestations of two different phenomena and hence, it is not appropriate to generate $\mathbf{x}$ by simultaneously minimizing $\frac{1}{p}\|\mathbf{x}-\mathbf{u}\|_p^p$ and $\frac{1}{q}\|\mathbf{x}-\mathbf{v}\|_q^q$. 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

$$\mathcal{E}_2(\mathbf{x}) = \frac{1}{q}\|\nabla\mathbf{x}-\nabla\mathbf{v}\|_q^q, \tag{2}$$

where ∇ is the gradient operator which we will define in details latter. In the case of $q=0$, Eq. (2) is defined as $\mathcal{E}_2(\mathbf{x}) = \|\nabla\mathbf{x}-\nabla\mathbf{v}\|_0$, which equals the number of non-zero entries of $\nabla\mathbf{x}-\nabla\mathbf{v}$. Combining Eqs. (1) and (2), we formulate the fusion problem as minimizing the following objective function:

$$\begin{aligned}\mathcal{E}(\mathbf{x}) &= \mathcal{E}_1(\mathbf{x}) + \lambda\mathcal{E}_2(\mathbf{x})\\ &= \frac{1}{p}\|\mathbf{x}-\mathbf{u}\|_p^p + \lambda\frac{1}{q}\|\nabla\mathbf{x}-\nabla\mathbf{v}\|_q^q,\end{aligned} \tag{3}$$

where the first term constrains the fused image $\mathbf{x}$ to have the similar pixel intensities with the infrared image $\mathbf{u}$, the second term requires that the fused image $\mathbf{x}$ and the visible image $\mathbf{v}$ have the similar gradients, more specifically, the similar edges in corresponding positions, and λ is a positive parameter controlling the trade-off between the two terms.

The objective function (3) to some extent aims to transfer the gradients/edges in the visible image onto the corresponding positions in the infrared image. Thus the fused image should still look like an infrared image, but with more appearance details, i.e., an infrared image with more complex and detailed scene representation. It plays a role of infrared image sharpness or enhancement, which is the major difference between our method and other typical fusion methods [30]. As our method fuses two images based on gradient transfer, we name our method Gradient Transfer Fusion (GTF).

2.2. Optimization using total variation minimization

Now we consider the ℓ^p and ℓ^q norms in the objective function (3). If the difference between the fused image $\mathbf{x}$ and the infrared image $\mathbf{u}$ is Gaussian then $p=2$ is the natural choice whereas $p=1$ is the better choice for Laplacian or impulsive case. Specifically, in our problem we expect to keep the thermal radiation information of $\mathbf{u}$, which means that most entries of $\mathbf{x}-\mathbf{u}$ should be zero while a small part of the entries could be large due to the purpose of gradient transfer from the visible image $\mathbf{v}$. Thus the difference between $\mathbf{x}$ and $\mathbf{u}$ should be Laplacian or impulsive rather than Gaussian, i.e. $p=1$. On the other hand, as natural images are often piece-wise smooth, their gradients tend to be sparse and gradients with large magnitude correspond to the edges. To encourage sparseness of the gradients, it mathematically amounts to minimizing the ℓ^0 norm, i.e., $q=0$. However, the ℓ^0 norm is NP-hard, and an alternative convex relaxation approach is to replace ℓ^0 by ℓ^1. The restricted isometry property condition [31] theoretically guarantees the exact recovery of sparse solutions by ℓ^1. Therefore, we consider minimizing the gradient differences with ℓ^1 norm, i.e. $q=1$, and ℓ^1 on the gradient is the total variation [22]. Let $\mathbf{y}=\mathbf{x}-\mathbf{v}$, the optimization problem (3) can be rewritten as:

$$\mathbf{y}^* = \arg\min_{\mathbf{y}}\left\{\sum_{i=1}^{mn}|\mathbf{y}_i-(\mathbf{u}_i-\mathbf{v}_i)| + \lambda J(\mathbf{y})\right\}, \tag{4}$$

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

where $|x| := \sqrt{x_1^2+x_2^2}$ for every $x=(x_1,x_2)\in\mathbb{R}^2$, $\nabla_i = (\nabla_i^h, \nabla_i^v)$ 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, $\nabla_i^h \mathbf{x} = \mathbf{x}_i - \mathbf{x}_{r(i)}$ and $\nabla_i^v \mathbf{x} = \mathbf{x}_i - \mathbf{x}_{b(i)}$, 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. Clearly, the objective function (4) is convex and thus has a global optimal solution. Its first term could be seen as a data fidelity term, and the second term could be seen as a regularization item, which strikes an appropriate balance in preserving the thermal radiation information and the detailed appearance information in the two given images.

The problem (4) is a standard ℓ^1-TV minimization problem, which can be solved by directly using the algorithm proposed in [23]. In summary, our GTF algorithm is very simple yet efficient. With a regularization parameter λ, it merely requires to compute $\mathbf{y}$ by optimizing the problem (4) via an existing ℓ^1-TV minimization technique, and the global optimal solution of the fused image $\mathbf{x}^*$ is then determined by: $\mathbf{x}^* = \mathbf{y}^* + \mathbf{v}$.

3. Fusion of two unregistered images

Generally, successful image fusion requires an essential and challenging step that the image pairs to be fused have to be correctly co-registered on a pixel-by-pixel basis. The proposed GTF also requires the infrared/visible pairs to be aligned in advance. In this section, we generalize our GTF so that it is able to fuse unregistered image pairs.

3.1. The generalized GTF algorithm

We denote by $\mathcal{T}$ the spatial transformation between the infrared and visible image pair, i.e., $\mathbf{v}(\mathcal{T})$ and $\mathbf{u}$ should be co-registered ideally. The fusion is then performed on $\mathbf{u}$ and $\mathbf{v}(\mathcal{T})$. Note that in this problem the sizes of the original infrared and visible images are not required to be the same. As we choose $p = 1$ and $q = 1$, the objective function (3) becomes

$$\mathcal{E}(\mathbf{x}, \mathcal{T}) = \|\mathbf{x} - \mathbf{u}\|_1 + \lambda \|\nabla \mathbf{x} - \nabla \mathbf{v}(\mathcal{T})\|_1. \tag{5}$$

There are two unknown variables in the objective function (5), i.e., the fused image $\mathbf{x}$ and the transformation $\mathcal{T}$. The difficulty arises when the variable is being solved with no information at all on the other variable. By contrast, when the other variable is known, the procedure becomes simpler than that for the coupled problem. For example, the problem can be solved by iteratively solving $\mathbf{x}$ and $\mathcal{T}$ by fixing the other variable until convergence. To solve $\mathbf{x}$ by fixing $\mathcal{T}$, we let $\mathbf{y} = \mathbf{x} - \mathbf{v}(\mathcal{T})$ and obtain a similar total variation minimization problem as in Eq. (4), which can be solved according to our GTF algorithm:

$$\mathbf{y}^* = \arg\min_{\mathbf{y}} \left\{ \frac{1}{2} \|\mathbf{y} - (\mathbf{u} - \mathbf{v}(\mathcal{T}))\|_1 + \lambda J(\mathbf{y}) \right\}. \tag{6}$$

To solve $\mathcal{T}$ by fixing $\mathbf{x}$, the objective function (5) with respect to $\mathcal{T}$ can be written as:

$$\begin{aligned} \mathcal{E}(\mathcal{T}) &= \|\nabla \mathbf{x} - \nabla \mathbf{v}(\mathcal{T})\|_1 \\ &= \sum_{i=1}^{mn} \sqrt{(\nabla_i^h \mathbf{x} - \nabla_i^h \mathbf{v}(\mathcal{T}))^2 + (\nabla_i^v \mathbf{x} - \nabla_i^v \mathbf{v}(\mathcal{T}))^2}. \end{aligned} \tag{7}$$

The transformation $\mathcal{T}$ could be modeled by using rigid or non-rigid model [32]. As the infrared and visible images are typically captured simultaneously from the similar viewpoint, an affine model with six motion parameters will be accurate enough to approximate the transformation. Techniques involving gradient-based numerical optimization, such as the quasi-Newton method, are used here to iteratively update the transformation parameters $\mathcal{T}$. In computing the objective function and its gradient, the following pseudo-code is adopted

$$\mathbf{r} = (\nabla^h \mathbf{x} - \nabla^h \mathbf{v}(\mathcal{T}), \nabla^v \mathbf{x} - \nabla^v \mathbf{v}(\mathcal{T})), \tag{8}$$

$$\mathbf{s} = (\nabla^h \mathbf{v}(\mathcal{T}), \nabla^v \mathbf{v}(\mathcal{T})), \quad \mathcal{E} = \mathrm{tr}((\mathbf{r}\mathbf{r}^{\mathrm{T}})^{1/2}),$$

$$\nabla \mathcal{E} = -\frac{1}{2} (\mathbf{r}\mathbf{r}^{\mathrm{T}})^{-1/2} \mathbf{r} \cdot \nabla \mathbf{s} \cdot \frac{\partial \mathcal{T}}{\partial \theta},$$

where $\mathbf{r}$ and $\mathbf{s}$ are matrices of size $mn \times 2$, $\mathrm{tr}(\cdot)$ denotes the trace, $\nabla \mathbf{s}$ can be obtained from the gradients of the horizontal gradient image $\nabla^h \mathbf{v}(\mathcal{T})$ and vertical gradient image $\nabla^v \mathbf{v}(\mathcal{T})$, and θ represents the transformation parameters.

3.2. Transformation initialization

It is obvious that a good initialization for the transformation can significantly accelerate the convergence of the iteration. Thus we need to seek a solution of the transformation that can roughly align the image pair. Due to the different phenomena of the image pair, appearance features such as graylevels/colors, textures and gradient histograms are not likely to match. Instead, a major strategy that address this issue is to first extract salient structures in the image represented as compact geometrical entities (e.g., points of high curvature, line intersections, strong edges, structural contours and silhouettes), and then seek to achieve registration indirectly through the alignment of the geometrical features under a geometric constraint (e.g., affine transformation in this paper).

For the salient structure, we choose the edge map, which is a significant common feature that might be preserved in an infrared/visible image pair [2,33]. To this end, we choose the binarized edge maps extracted by the Canny edge detectors to represent the images [2]. After we obtain the edge maps, we discretize them into a set of points by using a sampling method introduced in [34]. The sampling method selects a subset of the edge pixels ensuring that the sample points have a certain minimum distance between them, as this makes sure that the sampling along the object contours is somewhat uniform. Thus the binary edge maps of the infrared and visible data can be described as two point sets and hence, point set registration methods such as [35,36] can be applied to estimate the transformation parameters. In this paper, the Coherent Point Drift (CPD) [35] algorithm is chosen which considers the alignment of two point sets as a probability density estimation problem by using Gaussian mixture models. We summarize the proposed method for fusion without pre-alignment in Algorithm 1.

Algorithm 1: Gradient Transfer Fusion without pre-alignment.

Input: Infrared image $\mathbf{u}$, visible image $\mathbf{v}$, parameter λ
Output: Fused image $\mathbf{x}$

1. Extract Canny edge maps of $\mathbf{u}$ and $\mathbf{v}$ and discretize them into point sets by using the sampling method [34];
2. Align the two point set by using CPD [35] with an affine model and initialize the transformation $\mathcal{T}$ accordingly;
3. **repeat**
4. Compute $\mathbf{y}$ by optimizing the problem (6) using the algorithm introduced in [22];
5. Set the fused image $\mathbf{x} \leftarrow \mathbf{y} + \mathbf{v}(\mathcal{T})$;
6. Compute $\mathcal{T}$ by optimizing the problem (7) using the pseudo-code (8) together with the quasi-Newton method;
7. **until** *Stop criterions*;
8. The final fused image $\mathbf{x}$ is calculated as $\mathbf{x} = \mathbf{y} + \mathbf{v}(\mathcal{T})$.

4. Experimental results

In this section, we test the performance of our GTF on publicly available datasets, and compare it with eight state-of-the-art fusion methods such as Laplacian pyramid (LP) [9], ratio of low-pass pyramid (RP) [11], Wavelet [12], dual-tree complex wavelet transform (DTCWT) [13], curvelet transform (CVT) [14], multi-resolution singular value decomposition (MSVD) [15], guided filtering based fusion (GFF) [16], and Laplacian pyramid with sparse representation (LP-SR) [18]. All the eight algorithms are implemented based on publicly available codes, where the parameters are set according to the original paper, and we try our best to tune some details. The experiments are performed on a laptop with 3.3 GHz Intel Core CPU, 8GB memory and Matlab Code.

4.1. Datasets and settings

In our experimental evaluation we first focus on qualitative and quantitative comparisons on the fusion performance of different methods on aligned infrared and visible image pairs, and then show results of our GTF algorithm on unaligned infrared and visible image pairs. Next, we discuss the datasets used in this paper.

- Aligned dataset: We test our method on the surveillance images from TNO Human Factors, which contains multispectral nighttime imagery of different military relevant scenarios, registered with different multiband camera systems[1]. We choose seven typical pairs with names *Bunker*, *Lake*, *Tank*, *Bench*, *Sandpath*, *Nato_camp* and *Dune* for qualitative illustration, while *Nato_camp* and *Dune* are two image sequences and are further used for quantitative comparison. The two sequences contain 32 and 23 image pairs, respectively. In addition, the source images are enhanced in terms of their representation using the procedure of contrast stretching.
- Unaligned dataset: The publicly available benchmark UTK-IRIS (Imaging, Robotics and Intelligent System) Thermal/Visible Face Database is used for evaluation[2]. It contains a number of individuals with various poses, facial expressions, as well as illumination changes. The images are all of resolution 320 × 240, and we choose several typical pairs for qualitative illustration.

The results of image fusion are typically assessed either subjectively or objectively. Since there is insignificant difference among the fusion results in most circumstances, difficulty arises from the accurate subjective evaluation. Recent studies have proposed various fusion metrics. The basis for most of these metrics is the measurement of the transfer of a feature, such as edges and amount of information, from the source images into the new fused composite image [37]. However, none of them is definitely better than others. Therefore, it is necessary to apply multiple metrics to make a comprehensive evaluation of different fusion methods. In this paper, we quantitatively evaluate the performances of different fusion methods using four metrics, i.e., mutual information (MI) [38], gradient based fusion metric (Q^G) [39], feature mutual information (FMI) [40], and entropy (EN) [18]. The definitions of these four metrics are as follows:

- Mutual information (MI): A foundational concept of information theory is MI, which measures how much information a given variable contains about another variable [38]. This concept is employed to assess the performance of image fusion. In detail, MI measures the extent at which two variables, $\mathbf{u}$ and $\mathbf{v}$, depend on each other. Such dependency can be determined, by calculating through relative entropy, the distance between the joint distribution $P_{\mathbf{uv}}(u,v)$ and the distribution related to complete independence $P_{\mathbf{u}}(u)\cdot P_{\mathbf{v}}(v)$. The joint and marginal distributions, $P_{\mathbf{uv}}(u,v)$, $P_{\mathbf{u}}(\mathbf{u})$ and $P_{\mathbf{v}}(v)$, are simply obtained by normalizing the joint and marginal histograms of both images, respectively.
- Gradient based fusion metric (Q^G): The Q^G is a popular fusion metric, which evaluates the amount of gradient information transferred into the fused image from the infrared and visible images. Q^G is calculated as follows:

$$Q^G(\mathbf{x}) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(Q^{\mathbf{ux}}(i,j)W^{\mathbf{u}}(i,j)+Q^{\mathbf{vx}}(i,j)W^{\mathbf{v}}(i,j))}{\sum_{i=1}^{m}\sum_{j=1}^{n}(W^{\mathbf{u}}(i,j)+W^{\mathbf{v}}(i,j))}, \tag{9}$$

  where $Q^{\mathbf{ux}}(i,j)=Q_g^{\mathbf{ux}}(i,j)Q_\alpha^{\mathbf{ux}}(i,j)$, $Q_g^{\mathbf{ux}}(i,j)$ and $Q_\alpha^{\mathbf{ux}}(i,j)$ represent the edge strength and orientation preservation values at pixel location (i,j), respectively, $W^{\mathbf{u}}(i,j)$ indicates the importance of $Q^{\mathbf{ux}}(i,j)$, and similarly for the definitions of $Q^{\mathbf{vx}}(i,j)$ and $W^{\mathbf{v}}(i,j)$.
- FMI measures the mutual information found in all the image features. The most appropriate feature is chosen, in case of any specific application, to make this method adaptable [40]. The gradient map is one of the most convenient and frequently used features of an image. This map contains information on the following: contrast, texture, pixel neighborhoods, and edge strength and directions. From the source and fused images, the distribution functions of the above features can be directly extracted, and the marginal distributions consist of the normalized values of the gradient magnitude feature images.
- Entropy (EN): The EN as a measure of the amount of information of the fused image is defined as follows:

$$\mathrm{EN}(\mathbf{x}) = -\sum_{l=0}^{L-1} P_{\mathbf{x}}(l)\log_2 P_{\mathbf{x}}(l), \tag{10}$$

  where L is the number of gray level, which is set to 256 in our experiments, and $P_{\mathbf{x}}(l)$ is the normalized histogram of the fused image $\mathbf{x}$.

For all the four metrics, larger values indicate better performance.

Parameter initialization. There is only one parameter in our GTF algorithm, i.e., the regularization parameter λ. It controls the trade-off of the information preservation between the two source images. Generally, small value of λ indicates keeping more thermal radiation information, while large value of λ corresponds to transferring more texture information. The fusion result will be the original infrared image when $\lambda = 0$ and the original visible image when $\lambda = +\infty$. Throughout our experiments, we fix λ to 4 as an empirical value, which can achieve good visual effects in most cases.

4.2. Results on aligned dataset

To get an intuitive impression of the proposed GTF's performance, our first experiment involves infrared and visible image fusion on seven typical aligned pairs, i.e. *Bunker*, *Lake*, *Tank*, *Bench*, *Sandpath*, *Nato_camp* and *Dune*, as shown in Fig. 2. In the *Lake* pair, the lake surface is dark in the visible image and bright in the infrared image, while the waterweeds are bright in the visible image and dark in the infrared image. In the rest pairs, the backgrounds (such as the woods) are clear in the visible images, while the targets (i.e., building, tank and humans) are salient in the infrared images.

The qualitative comparisons of our GTF and other eight methods such as LP [9], RP [11], Wavelet [12], DTCWT [13], CVT [14],

[1] The dataset is available at http://figshare.com/articles/TNO_Image_Fusion_Dataset/1008029.

[2] The dataset is available at http://www.cse.ohio-state.edu/otcbvs-bench/.

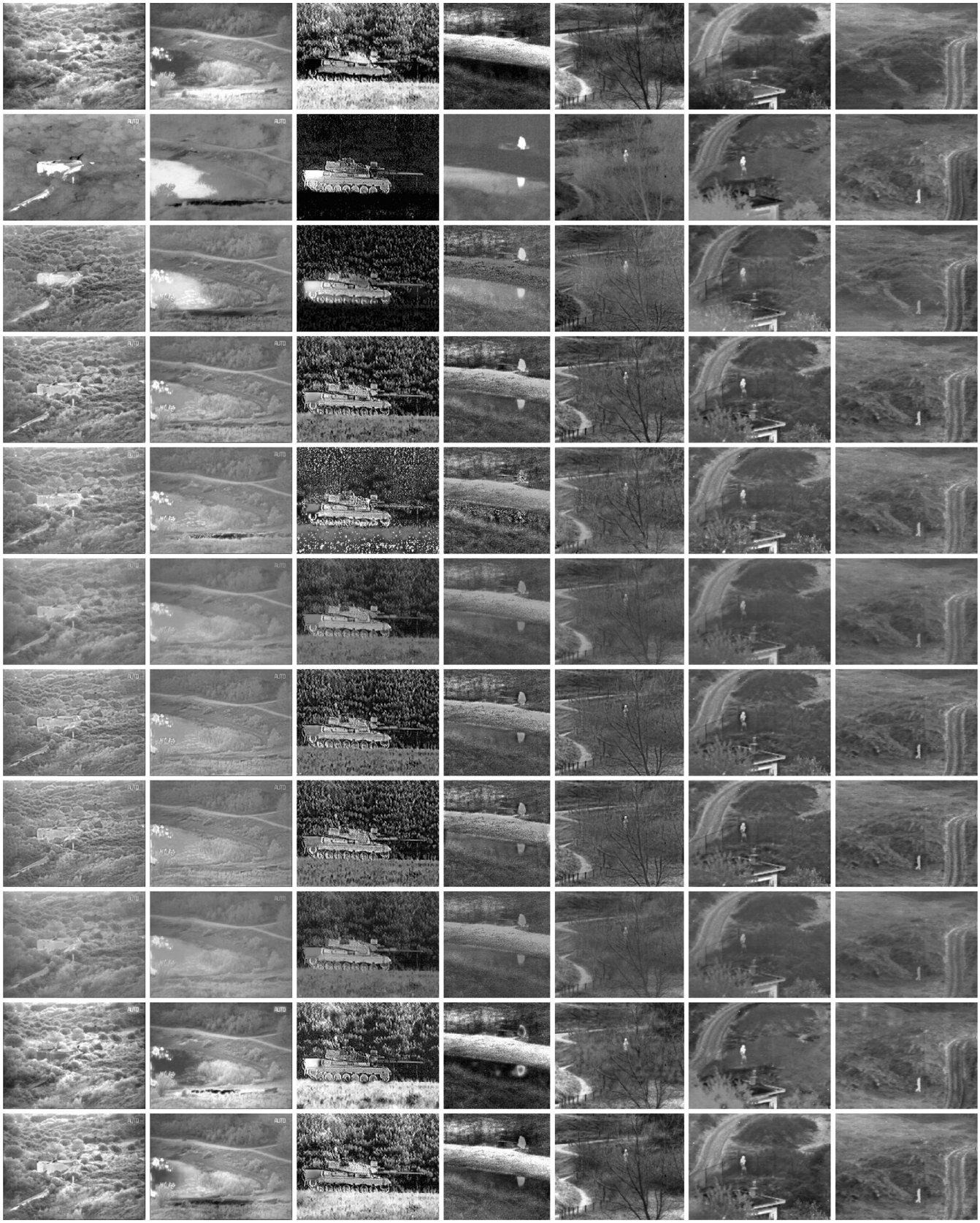

Fig. 2. Qualitative fusion results on the seven image pairs of *Bunker*, *Lake*, *Tank*, *Bench*, *Sandpath*, *Nato_camp* and *Dune* (from left to right). From top to bottom: visible images, infrared images, our fusion results, results of LP [9], RP [11], Wavelet [12], and DTCWT [13], CVT [14], MSVD [15], GFF [16], and LP-SR [18], respectively. Clearly, only our method can preserve both the thermal radiation and the texture information in the source images.

Fig. 3. Qualitative fusion results on the seven image pairs of *Bunker*, *Lake*, *Tank*, *Bench*, *Sandpath*, *Nato_camp* and *Dune* (from left to right). From top to bottom: infrared images, our fusion results when $\lambda = 0.1, 1, 3, 5, 7, 9, 100$ and visible images, respectively.

MSVD [15], GFF [16] and LP-SR [18] are given in Fig. 2. For each group of results, the first three rows present the original visible image, the original infrared image, and the fusion result of our GTF, respectively; while the rest eight rows correspond to the fusion results of the other eight methods. From the results, we can see that for the purpose of keeping details of scenes, all the nine methods work well, as the details of the waterweeds in the *Lake* pair, the details of the woods and the contours of the targets in all the other pairs are well preserved in the fused images. In this sense, it is hard to judge which method is the best or worst. However, for the compared methods, the pixel intensities of the fused images are to some extent the average of the two source images, which may lead to that the target is not well highlighted, as can be seen from the regions of lake surface and targets such as building, tank and humans. In contrast, the thermal radiation information in our results is well preserved, which makes the fused images look like high-resolution infrared images with clear highlighted targets. This will be beneficial for automatic target detection and localization, especially in the military applications. It is also the major difference between our method and the other existing fusion methods.

An advantage of our method is that parameter λ can modulate the information and quality of a fused image. To demonstrate this property, we provide qualitative fusion results of different λ on the seven image pairs (e.g., *Bunker*, *Lake*, *Tank*, *Bench*, *Sandpath*, *Nato_camp* and *Duine*), as shown in Fig. 3. From top to bottom are respectively the infrared images, our fusion results when λ =0.1, 1, 3, 5, 7, 9, 100, and the visible images. We see that as the value of λ increases, the fused image transforms from the infrared image to the visible image, and the fusion result is satisfied when λ is set to be around 4. Note that some bright small-scale details such as the "AUTO" in the *Lake* pair disappear in the fused image. This is due to that they only exist in the infrared images, and do not exist in the visible images. As our formulation attempts to preserve the thermal radiation information in the infrared image as well as the detailed appearance information (the edges) in the visible images, these small-scale details which only exist in the infrared image will be removed in the fused image. Nevertheless, the parameter λ can control the trade-off between preserving the thermal radiation information and the detailed appearance information. With a smaller value of λ, the formulation attempts to preserve more information in the infrared image and hence, the small-scale details existing only in the infrared images can be preserved. This can be seen from the *Lake* pair, where the "AUTO" is preserved very well when $\lambda < 1$.

We next give quantitative comparisons of the nine methods on two infrared/visible image sequences, such as *Nato_camp* and *Dune*. Some examples of image pairs from the two sequences and the corresponding fusion results are shown in the last two columns in Fig. 2. From the results, we can see how information from infrared and visible images improves situational awareness in a military

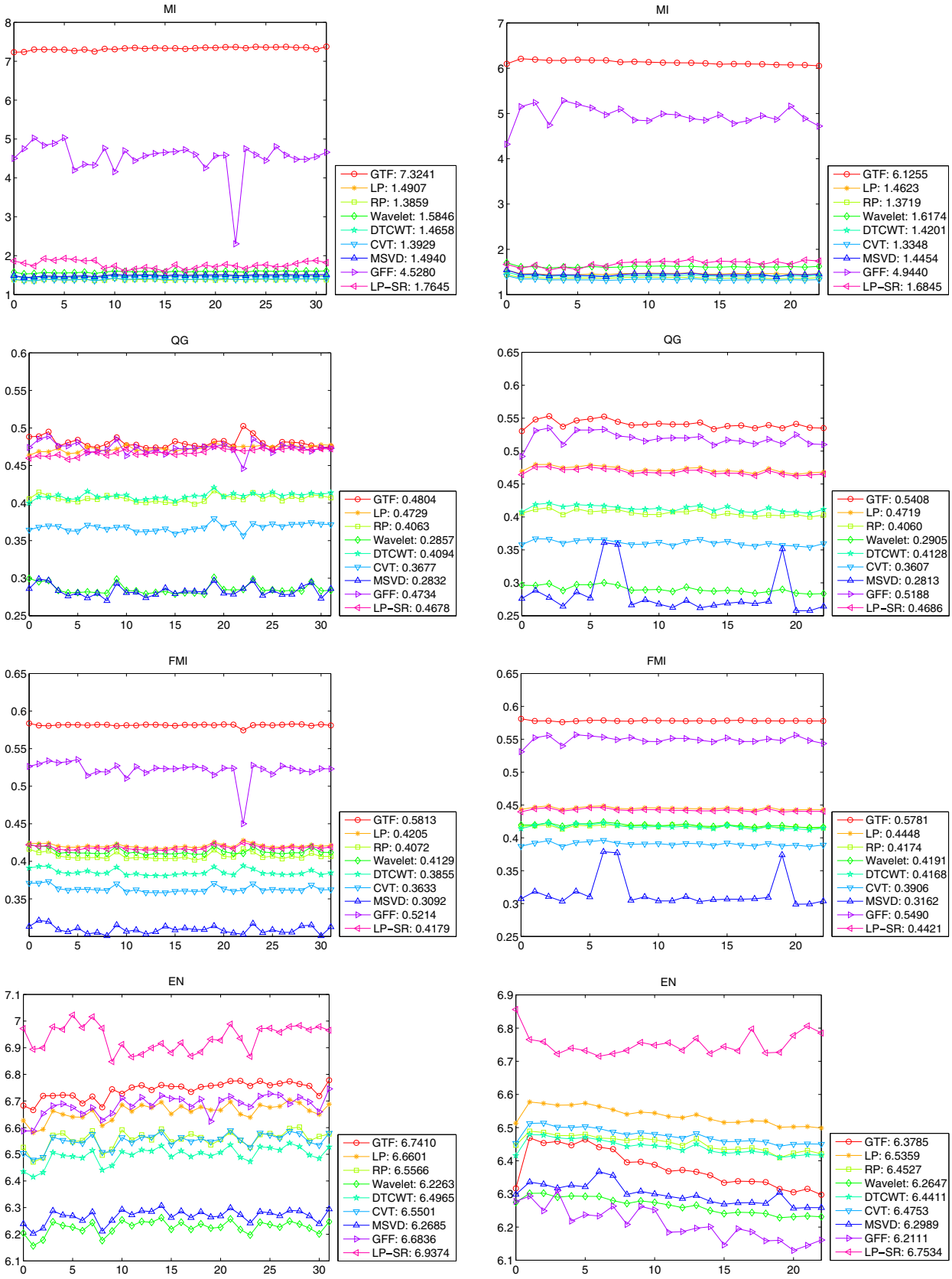

Fig. 4. Quantitative comparisons of the four metrics, i.e., MI, Q^G, FMI and EN on the *Nato_camp* (left column) and *Dune* (right column) sequences. The eight state-of-the-art methods such as LP [9], RP [11], Wavelet [12], DTCWT [13], CVT [14], MSVD [15], GFF [16] and LP-SR [18] are used for comparison. For all the four metrics, larger values indicate better performance.

campus. In a visible image, the process of distinguishing a person in camouflage from his background is difficult, however, this person gains in clarity and becomes more distinct in an IR image. In the visible image, however, the easily distinguishable background becomes nearly imperceptible in the IR image; an example is the fence in the *Nato_camp* sequence. Both images can be combined in our fused image into a distinct composite that provides an enhanced rendition of the complete scene relative to that provided by either of the two original images.

The quantitative comparisons on the two sequences are given in Fig. 4. We see that the curves are similar on the two sequences. Our GTF marked with red circles consistently has the best MIs and FMIs, and has the best Q^G on most image pairs, followed by the GFF and LP-SR methods. For the metric of EN, our GTF can also yield comparable results, especially on the *Nato_camp* sequence, where it can consistently achieve the second best performance. It is interesting that the traditional LP method performs better than several other recent state-of-the-art methods on both sequences. The run time comparison of nine algorithms on the two sequences are given in Table 1, where the images are all of size 270 × 360, and each value denotes the mean and variance of run time of a certain method on a sequence. We see that our method can achieve comparable efficiency compared to the other eight methods.

From the results, we can draw a conclusion that fusing thermal radiation and texture information as in our formulation can not only identify the most important information, but also can keep the largest or approximately the largest amount of information in the source images compared to existing state-of-the-art fusion methods.

Table 1
Run time comparison of nine algorithms on the *Nato_camp* and *Dune* sequences, where each value denotes the mean and variance of run time of a certain method on a sequence (unit: second).

Method	*Nato_camp*	*Dune*
GTF	$0.13 \pm 2.24 \times 10^{-4}$	$0.13 \pm 2.42 \times 10^{-4}$
LP [9]	$4.50 \times 10^{-3} \pm 1.50 \times 10^{-7}$	$4.43 \times 10^{-3} \pm 2.53 \times 10^{-8}$
RP [11]	$5.13 \times 10^{-2} \pm 8.97 \times 10^{-7}$	$5.13 \times 10^{-2} \pm 4.27 \times 10^{-7}$
Wavelet [12]	$0.14 \pm 3.45 \times 10^{-6}$	$0.14 \pm 2.13 \times 10^{-6}$
DTCWT [13]	$0.16 \pm 1.19 \times 10^{-5}$	$0.16 \pm 3.00 \times 10^{-6}$
CVT [14]	$1.10 \pm 4.69 \times 10^{-3}$	$1.07 \pm 1.65 \times 10^{-4}$
MSVD [15]	$0.20 \pm 8.76 \times 10^{-5}$	$0.20 \pm 1.59 \times 10^{-5}$
GFF [16]	$0.10 \pm 1.05 \times 10^{-4}$	$0.10 \pm 8.09 \times 10^{-6}$
LP-SR [18]	$1.61 \times 10^{-2} \pm 5.46 \times 10^{-7}$	$1.37 \times 10^{-2} \pm 1.47 \times 10^{-7}$

4.3. Results on unaligned dataset

As our GTF is able to fuse unaligned image pairs, in this section we further test its fusion as well as registration performance on the face dataset of UTK-IRIS. For the face recognition problem, fusing infrared and visible frames has been shown to be able to

Fig. 5. Fusion results of our GTF on six typical unregistered visible/IR image pairs in the dataset of UTK-IRIS, which involves different poses, facial expressions, and illumination changes. From left to right: original visible images, original infrared images, fused images, aligned visible images, and aligned edge maps. In the last column, the blue pluses denote Canny edge maps of the original infrared images, and the red circles indicate Canny edge maps of the original visible images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

achieve higher recognition rate [5]. Since existing fusion methods (including all the eight methods used for comparison) typically operate on aligned image pairs, we only qualitatively demonstrate our results on several typical pairs, as shown in Fig. 5.

In Fig. 5, the first two rows are image pairs in bright light, the middle two rows are image pairs in dark conditions, and the last two rows are image pairs with different facial expressions. The visible, infrared and fused images are given in the first, second and third columns, respectively. We see that the fused images look like sharpened infrared images. Compared to the infrared images, they contain much more details, which can be seen from the eyebrow, eyes behind the glasses, noses, ears, mouths, etc. While compared to the visible images, the fused images involve the robust thermal radiation information, which is beneficial in case of illumination changes, such as the first four rows. In our evaluation, we align the visible images to the infrared images, and the transformed visible images are presented in the fourth column. Besides, to observe the registration performance in details, we also provide the Canny edge map alignment results in the last column. We see that our GTF can generate almost perfect alignments on all the six face pairs, as the common edges such as the silhouettes of the faces and the eyeglasses are accurately registered.

5. Conclusion

Within this paper, we propose a novel infrared and visible fusion method called *Gradient Transfer Fusion* (GTF) based on gradient transfer and total variation minimization. It can simultaneously keep the thermal radiation information in the infrared image and preserve appearance information in the visible image. The fusion results look like high-resolution infrared images with clear highlighted targets and hence, it will be beneficial for fusion-based target detection and recognition systems. To enhance the applicability of the proposed method, we also generalize the formulation so that it is able to fuse image pairs without pre-registration. The quantitative comparisons on several metrics with other eight state-of-the-art fusion methods demonstrate that our method can not only identify the most important information, but also can keep the largest or approximately the largest amount of information in the source images.

Acknowledgments

The authors gratefully acknowledge the financial supports from the National Natural Science Foundation of China under Grant nos. 61503288 and 41501505, and the China Postdoctoral Science Foundation under Grant nos. 2015M570665 and 2015M572194.

References

[1] H. Li, B. Manjunath, S.K. Mitra, Multisensor image fusion using the wavelet transform, Graph. Model. Image Process. 57 (1995) 235–245.
[2] J. Ma, J. Zhao, Y. Ma, J. Tian, Non-rigid visible and infrared face registration via regularized gaussian fields criterion, Pattern Recognit. 48 (2015) 772–784.
[3] C. Chen, Y. Li, W. Liu, J. Huang, Image fusion with local spectral consistency and dynamic gradient sparsity, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, 2014, pp. 2760–2765.
[4] A. Toet, J.K. IJspeert, A.M. Waxman, M. Aguilar, Fusion of visible and thermal imagery improves situational awareness, Displays 18 (1997) 85–95.
[5] S.G. Kong, J. Heo, F. Boughorbel, Y. Zheng, B.R. Abidi, A. Koschan, M. Yi, M.A. Abidi, Multiscale fusion of visible and thermal ir images for illumination-invariant face recognition, Int J. Comput. Vis. 71 (2007) 215–233.
[6] G. Piella, A general framework for multiresolution image fusion: from pixels to regions, Inf. Fusion 4 (2003) 259–280.
[7] J. Saeedi, K. Faez, Infrared and visible image fusion using fuzzy logic and population-based optimization, Appl. Soft Comput. 12 (2012) 1041–1054.
[8] V.S. Petrovic, C.S. Xydeas, Gradient-based multiresolution image fusion, IEEE Trans. Image Process. 13 (2004) 228–237.
[9] P.J. Burt, E.H. Adelson, The laplacian pyramid as a compact image code, IEEE Trans. Commun. 31 (1983) 532–540.
[10] A. Toet, L.J. Van Ruyven, J.M. Valeton, Merging thermal and visual images by a contrast pyramid, Opt. Eng. 28 (1989) 287789.
[11] A. Toet, Image fusion by a ratio of low-pass pyramid, Pattern Recognit. Lett. 9 (1989) 245–253.
[12] P. Zeeuw, Wavelets and image fusion, CWI, Amsterdam, March, 1998.
[13] J.J. Lewis, R.J. O'Callaghan, S.G. Nikolov, D.R. Bull, N. Canagarajah, Pixel-and region-based image fusion with complex wavelets, Inf. Fusion 8 (2007) 119–130.
[14] F. Nencini, A. Garzelli, S. Baronti, L. Alparone, Remote sensing image fusion using the curvelet transform, Inf. Fusion 8 (2007) 143–156.
[15] V. Naidu, Image fusion technique using multi-resolution singular value decomposition, Def. Sci. J. 61 (2011) 479–484.
[16] S. Li, X. Kang, J. Hu, Image fusion with guided filtering, IEEE Trans. Image Process. 22 (2013a) 2864–2875.
[17] S. Li, X. Kang, J. Hu, B. Yang, Image matting for fusion of multi-focus images in dynamic scenes, Inf. Fusion 14 (2013b) 147–162.
[18] 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.
[19] B. Yang, S. Li, Pixel-level image fusion with simultaneous orthogonal matching pursuit, Inf. Fusion 13 (2012) 10–19.
[20] G. Pajares, J.M. De La Cruz, A wavelet-based image fusion tutorial, Pattern Recognit. 37 (2004) 1855–1872.
[21] A.L. Da Cunha, J. Zhou, M.N. Do, The nonsubsampled contourlet transform: theory, design, and applications, IEEE Trans. Image Process. 15 (2006) 3089–3101.
[22] A. Chambolle, An algorithm for total variation minimization and applications, J. Math. Imaging Vis. 20 (1-2) (2004) 89–97.
[23] T.F. Chan, S. Esedoglu, Aspects of total variation regularized L^1 function approximation, SIAM J. Appl. Math. 65 (2005) 1817–1837.
[24] R.N. Bracewell, R. Bracewell, The Fourier transform and its applications, McGraw-Hill, New York, 1986.
[25] P. Viola, W.M. Wells III, Alignment by maximization of mutual information, Int. J. Comput. Vis. 24 (1997) 137–154.
[26] Y. Liu, Improving ICP with easy implementation for free-form surface matching, Pattern Recognit. 37 (2004) 211–226.
[27] J. Ma, J. Zhao, A.L. Yuille, Non-rigid point set registration by preserving global and local structures, IEEE Trans. Image Process 25 (2016) 53–64.
[28] J. Ma, H. Zhou, J. Zhao, Y. Gao, J. Jiang, J. Tian, Robust feature matching for remote sensing image registration via locally linear transforming, IEEE Trans. Geosci. Remote Sens. 53 (2015) 6469–6481.
[29] B. Zitová, J. Flusser, Image registration methods: a survey, Image Vis. Comput. 21 (2003) 977–1000.
[30] B. Khaleghi, A. Khamis, F.O. Karray, S.N. Razavi, Multisensor data fusion: a review of the state-of-the-art, Inf. Fusion 14 (2013) 28–44.
[31] E.J. Candes, J.K. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Commun. Pure Appl. Math. 59 (2006) 1207–1223.
[32] J. Ma, J. Zhao, J. Tian, A.L. Yuille, Z. Tu, Robust point matching via vector field consensus, IEEE Trans. Image Process. 23 (2014) 1706–1721.
[33] E. Coiras, J. Santamarı, C. Miravet, et al., Segment-based registration technique for visual-infrared images, Opt. Eng. 39 (2000) 282–289.
[34] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 509–522.
[35] A. Myronenko, X. Song, Point set registration: coherent point drift, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 2262–2275.
[36] J. Ma, W. Qiu, J. Zhao, Y. Ma, A.L. Yuille, Z. Tu, Robust L_2E estimation of transformation for non-rigid registration, IEEE Trans. Signal Process. 63 (2015) 1115–1129.
[37] N. Cvejic, T. Seppanen, S.J. Godsill, A non reference image fusion metric based on the regional importance measure, IEEE J. Sel. Top. Signal Process. 3 (2009) 212–221.
[38] G. Qu, D. Zhang, P. Yan, Information measure for performance of image fusion, Electron. Lett. 38 (2002) 313–315.
[39] C. Xydeas, V. Petrović, Objective image fusion performance measure, Electron. Lett. 36 (2000) 308–309.
[40] M.B.A. Haghighat, A. Aghagolzadeh, H. Seyedarabi, A non-reference image fusion metric based on mutual information of image features, Comput. Electr. Eng. 37 (2011) 744–756.