Iterative guided image fusion

Alexander Toet
TNO Soesterberg, Netherlands

ABSTRACT

We propose a multi-scale image fusion scheme based on guided filtering. Guided filtering can effectively reduce noise while preserving detail boundaries. When applied in an iterative mode, guided filtering selectively eliminates small scale details while restoring larger scale edges. The proposed multi-scale image fusion scheme achieves spatial consistency by using guided filtering both at the decomposition and at the recombination stage of the multi-scale fusion process. First, size-selective iterative guided filtering is applied to decompose the source images into approximation and residual layers at multiple spatial scales. Then, frequency-tuned filtering is used to compute saliency maps at successive spatial scales. Next, at each spatial scale binary weighting maps are obtained as the pixelwise maximum of corresponding source saliency maps. Guided filtering of the binary weighting maps with their corresponding source images as guidance images serves to reduce noise and to restore spatial consistency. The final fused image is obtained as the weighted recombination of the individual residual layers and the mean of the approximation layers at the coarsest spatial scale. Application to multiband visual (intensified) and thermal infrared imagery demonstrates that the proposed method obtains state-of-the-art performance for the fusion of multispectral nightvision images. The method has a simple implementation and is computationally efficient.

INTRODUCTION

The increasing deployment and availability of co-registered multimodal imagery from different types of sensors has spurred the development of image fusion techniques. The information provided by different sensors registering the same scene can either be (partially) redundant or complementary and may be corrupted with noise. Effective combinations of complementary and partially redundant multispectral imagery can therefore visualize information that is not directly evident from the individual input images. For instance, in nighttime (low-light) outdoor surveillance applications, intensified visual (II) or near-infrared (NIR) imagery often provides a detailed but noisy representation of a scene. While different types of noise may result from several processes associated with the underlying sensor physics, additive noise is typically the predominant noise component encountered in II and NIR imagery (Petrovic & Xydeas, 2003). Additive noise can be modelled as a random signal that is simply added to the original signal. As a result, additive noise may obscure or distort relevant image details. In addition, targets of interest like persons or cars...
are sometimes hard to distinguish in II or NIR imagery because of their low luminance contrast. While thermal infrared (IR) imagery typically represents these targets with high contrast, their background (context) is often washed out due to low thermal contrast. In this case, a fused image that clearly represents both the targets and their background enables a user to assess the location of targets relative to landmarks in their surroundings, thus providing more information than either of the input images alone.

Some potential benefits of image fusion are: wider spatial and temporal coverage, decreased uncertainty, improved reliability, and increased system robustness. Image fusion has important applications in defense and security for situational awareness (Toet et al., 1997), surveillance (Shah et al., 2013; Zhu & Huang, 2007), target tracking (Motamed, Lherbier & Hamad, 2005; Zou & Bhanu, 2005), intelligence gathering (O’Brien & Irvine, 2004), concealed weapon detection (Bhatnagar & Wu, 2011; Liu et al., 2006; Toet, 2003; Xue & Blum, 2003; Xue, Blum & Li, 2002; Yajie & Mowu, 2009), detection of abandoned packages (Beyan, Yigit & Temizel, 2011) and buried explosives (Lepley & Averill, 2011), and face recognition (Kong et al., 2007; Singh, Vatsa & Noore, 2008). Other important image fusion applications are found in industry (Tian et al., 2009), art analysis (Zitová, Beneš & Blažek, 2011), agriculture (Bulanona, Burks & Alchanatis, 2009), remote sensing (Ghassemian, 2001; Jacobson & Gupta, 2005; Jacobson, Gupta & Cole, 2007; Jiang et al., 2011) and medicine (Agarwal & Bedi, 2015; Biswas, Chakrabarti & Dey, 2015; Daneshvar & Ghassemian, 2010; Singh & Khare, 2014; Wang, Li & Tian, 2014; Yang & Liu, 2013) (for a survey of different applications of image fusion techniques see Blum & Liu (2006).

In general, image fusion aims to represent the visual information from any number of input images in a single composite (fused) image that is more informative than each of the input images alone, eliminating noise in the process while preventing both the loss of essential information and the introduction of artefacts. This requires the availability of filters that combine the extraction of relevant image details with noise reduction.

To date, a variety of image fusion algorithms have been proposed. A popular class of algorithms are the multi-scale image fusion schemes, which decompose the source images into spatial primitives at multiple spatial scales, then integrate these primitives to form a new (‘fused’) multi-scale representation, and finally apply an inverse multi-scale transform to reconstruct the fused image. Examples of this approach are for instance the Laplacian pyramid (Burt & Adelson, 1983), the Ratio of Low-Pass pyramid (Toet, 1989b), the contrast pyramid (Toet, Van Ruyven & Valeton, 1989), the filter-subtract-decimate Laplacian pyramid (Burt, 1988; Burt & Kolczynski, 1993), the gradient pyramid (Burt, 1992; Burt & Kolczynski, 1993), the morphological pyramid (Toet, 1989a), the discrete wavelet transform (Lemeshevsky, 1999; Li, Manjunath & Mitra, 1993; Li, Kwok & Wang, 2002; Scheunders & De Backer, 2001), the shift invariant discrete wavelet transform (Lemeshevsky, 1999; Rockinger, 1997; Rockinger, 1999; Rockinger & Fechner, 1998), the contourlet (Yang et al., 2010), the shift-invariant shearlet transform (Wang, Li & Tian, 2014), the non-subsampled shearlet transform (Kong, Wang & Lei, 2015; Liu et al., 2016; Zhang et al., 2015), the ridgelet transform (Tao, Junping & Ye, 2005). The filters applied in several of the earlier techniques typically produce halo artefacts near edges. More recent methods like
shearlets, contourlets and ridgelets are better capable to preserve local image features but are often complex or time-consuming.

Non-linear edge-preserving smoothing filters such as anisotropic diffusion (Perona & Malik, 1990), robust smoothing (Black et al., 1998) and the bilateral filter (Tomasi & Manduchi, 1998) may appear effective tools to prevent artefacts that arise from spatial inconsistencies in multi-scale image fusion schemes. However, anisotropic diffusion tends to over-sharpen edges and is computationally expensive, which makes it less suitable for application in multi-scale fusion schemes (Farbman et al., 2008). The non-linear bilateral filter (BLF) assigns each pixel a weighted mean of its neighbors, with the weights decreasing both with spatial distance and with difference in value (Tomasi & Manduchi, 1998). While the BLF is quite effective at smoothing small intensity changes while preserving strong edges and has efficient implementations, it also tends to blur across edges at larger spatial scales, thereby limiting its value for application in multi-scale image decomposition schemes (Farbman et al., 2008). In addition, the BLF has the undesirable property that it can reverse the intensity gradient near sharp edges (the weighted average becomes unstable when a pixel has only few similar pixels in its neighborhood: He, Sun & Tang, 2013). In the joint (or cross) bilateral filter (JBLF) a second or guidance image serves to steer the edge stopping range filter thus preventing over- or under-blur near edges (Petschnigg et al., 2004). Zhang et al. (2014) showed that the application of the JBLF in an iterative framework results in size selective filtering of small scale details combined with the recovery of larger scale edges. The recently introduced Guided Filter (GF: He, Sun & Tang, 2013) is a computationally efficient, edge-preserving translation-variant operator based on a local linear model which avoids the drawbacks of bilateral filtering and other previous approaches. When the input image also serves as the guidance image, the GF behaves like the edge preserving BLF. Hence, the GF can gracefully eliminate small details while recovering larger scale edges when applied in an iterative framework.

In this paper we propose a multi-scale image fusion scheme, where iterative guided filtering is used to decompose the input images into approximate and residual layers at successive spatial scales, and guided filtering is used to construct the weight maps used in the recombination process.

The rest of this paper is organized as follows. ‘Edge Preserving Filtering’ briefly discusses the principles of edge preserving filtering and introduces (iterative) guided filtering. In ‘Related Work’ we discuss related work. ‘Proposed Method’ presents the proposed guided fusion based image fusion scheme. ‘Methods and Material’ presents the imagery and computational methods that were used to assess the performance of the new image fusion scheme. The results of the evaluation study are presented in ‘Results.’ Finally, in ‘Discussion and Conclusions’ the results are discussed and some conclusions are presented.

**EDGE PRESERVING FILTERING**

In this section we briefly introduce the edge preserving bilateral and joint bilateral filters, show how they are related to the guided filter, and how the application of a guided filter in an iterative framework results in size selective filtering of small scale image details combined with the recovery of larger scale edges.
Bilateral filter

Spatial filtering is a common operation in image processing that is typically used to reduce noise or eliminate small spurious details (e.g., texture). In spatial filtering the value of the filtered image at a given location is a function (e.g., a weighted average) of the original pixel values in a small neighborhood of the same location. Although low pass filtering or blurring (e.g., averaging with Gaussian kernel) can effectively reduce image noise, it also seriously degrades the articulation of (blurs) significant image edges. Therefore, edge preserving filters have been developed that reduce small image variations (noise or texture) while preserving large discontinuities (edges).

The bilateral filter is a non-linear filter that computes the output at each pixel as a Gaussian weighted average of their spatial and spectral distances. It prevents blurring across edges by assigning larger weights to pixels that are spatially close and have similar intensity values (Tomasi & Manduchi, 1998). It uses a combination of (typically Gaussian) spatial and a range (intensity) filter kernels that perform a blurring in the spatial domain weighted by the local variation in the intensity domain. It combines a classic low-pass filter with an edge-stopping function that attenuates the filter kernel weights at locations where the intensity difference between pixels is large. Bilateral filtering was developed as a fast alternative to the computationally expensive technique of anisotropic diffusion, which uses gradients of the filtering images itself to guide a diffusion process, avoiding edge blurring (Perona & Malik, 1990). More formally, at a given image location (pixel) $i$, the filtered output $O_i$ is given by:

$$
O_i = \frac{1}{K_i} \sum_{j \in \Omega} I_j f(\|i-j\|) g(\|I_i - I_j\|)
$$

(1)

where $f$ is the spatial filter kernel (e.g., a Gaussian centered at $i$), $g$ is the range or intensity (edge-stopping) filter kernel (centered at the image value at $i$), $\Omega$ is the spatial support of the kernel, and $K_i$ is a normalizing factor (the sum of the $f \cdot g$ filter weights).

Intensity edges are preserved since the bilateral filter decreases not only with the spatial distance but also with the intensity distance. Though the filter is efficient and effectively reduces noise while preserving edges in many situations, it has the undesirable property that it can reverse the intensity gradient near sharp edges (the weighted average becomes unstable when a pixel has only few similar pixels in its neighborhood: He, Sun & Tang, 2013).

In the joint (or cross) bilateral filter (JBLF) the range filter is applied to a second or guidance image $G$ (Petschnigg et al., 2004):

$$
O_i = \frac{1}{K_i} \sum_{j \in \Omega} I_j f(\|i-j\|) \cdot g(\|G_i - G_j\|).
$$

(2)

The JBLF can prevent over- or under- blur near edges by using a related image $G$ to guide the edge stopping behavior of the range filter. That is, the JBLF smooths the image $I$ while preserving edges that are also represented in the image $G$. The JBLF is particularly favored when the edges in the image that is to be filtered are unreliable (e.g., due to noise.
or distortions) and when a companion image with well-defined edges is available (e.g., in the case of flash/no-flash image pairs). Thus, in the case of filtering an II image for which a companion (registered) IR image is available, the guidance image may either be the II image itself or its IR counterpart.

**Guided filtering**

A guided image filter (He, Sun & Tang, 2013) is a translation-variant filter based on a local linear model. Guided image filtering involves an input image $I$, a guidance image $G$ and an output image $O$. The two filtering conditions are (i) that the local filter output is a linear transform of the guidance image $G$ and (ii) as similar as possible to the input image $I$. The first condition implies that

$$O_i = a_k G_i + b_k \quad \forall i \in \omega_k$$

where $\omega_k$ is a square window of size $(2r+1) \times (2r+1)$. The local linear model ensures that the output image $O$ has an edge only at locations where the guidance image $G$ has one, because $\nabla O = a \nabla G$. The linear coefficients $a_k$ and $b_k$ are constant in $\omega_k$. They can be estimated by minimizing the squared difference between the output image $O$ and the input image $I$ (the second filtering condition) in the window $\omega_k$, i.e., by minimizing the cost function $E$:

$$E(a_k, b_k) = \sum_{i \in \omega_k} \left( (a_k G_i + b_k - I_i)^2 + \epsilon a_k^2 \right)$$

where $\epsilon$ is a regularization parameter penalizing large $a_k$. The coefficients $a_k$ and $b_k$ can directly be solved by linear regression (He, Sun & Tang, 2013):

$$a_k = \frac{1}{|\omega|} \sum_{i \in \omega_k} G_i I_i - \overline{G_k} \overline{I_k} \sigma_k^2$$

$$b_k = \overline{I_k} - a_k \overline{G_k}$$

where $|\omega|$ is the number of pixels in $\omega_k$, $\overline{I_k}$ and $\overline{G_k}$ represent the means of respectively $I$ and $G$ over $\omega_k$, and $\sigma_k^2$ is the variance of $I$ over $\omega_k$.

Since pixel $i$ is contained in several different (overlapping) windows $\omega_k$, the value of $O_i$ in Eq. (3) depends on the window over which it is calculated. This can be accounted for by averaging over all possible values of $O_i$:

$$O_i = \frac{1}{|\omega|} \sum_{k | i \in \omega_k} (a_k G_k + b_k).$$

Since $\sum_{k | i \in \omega_k} a_k = \sum_{k | i \in \omega_k} a_k$ due to the symmetry of the box window Eq. (7) can be written as

$$O_i = \overline{a_i} G_i + \overline{b_i}$$

where $\overline{a_i} = \frac{1}{|\omega|} \sum_{k | i \in \omega_k} a_k$ and $\overline{b_i} = \frac{1}{|\omega|} \sum_{k | i \in \omega_k} b_k$ are the average coefficients of all windows overlapping $i$. Although the linear coefficients ($\overline{a_i}, \overline{b_i}$) vary spatially, their gradients will be
smaller than those of $G$ near strong edges (since they are the output of a mean filter). As a result we have $\nabla O \approx \nabla G$, meaning that abrupt intensity changes in the guiding image $G$ are still largely preserved in the output image $O$.

Equations (5), (6) and (8) define the guided filter. When the input image also serves as the guidance image, the guided filter behaves like the edge preserving bilateral filter, with the parameters $\varepsilon$ and the window size $r$ having the same effects as respectively the range and the spatial variances of the bilateral filter. Equations (8) can be rewritten as

$$O_i = \sum_j W_{ij}(G)I_j$$  \hspace{1cm} (9)

with the weighting kernel $W_{ij}$ depending only on the guidance image $G$:

$$W_{ij} = \frac{1}{|\omega|^2} \sum_{k: (i, j) \in \omega} \left( 1 + \frac{(G_i - \bar{G}_k)(G_j - \bar{G}_k)}{\sigma_k^2 + \varepsilon} \right).$$  \hspace{1cm} (10)

Since $\sum_j W_{ij}(G) = 1$ this kernel is already normalized.

The guided filter is a computationally efficient, edge-preserving operator which avoids the gradient reversal artefacts of the bilateral filter. The local linear condition formulated by Eq. (3) implies that its output is locally approximately a scaled version of the guidance image plus an offset. This makes it possible to use the guided filter to transfer structure from the guidance image $G$ to the output image $O$, even in areas where the input image $I$ is smooth (or flat). This structure-transferring filtering is an useful property of the guided filter, and can for instance be applied for feathering/matteing and dehazing (He, Sun & Tang, 2013).

**Iterative guided filtering**

Zhang et al. (2014) showed that the application of the joint bilateral filter (Eq. (2)) in an iterative framework results in size selective filtering of small scale details combined with the recovery of larger scale edges. In this scheme the result $G_{t+1}$ of the $t$th iteration is obtained from the joint bilateral filtering of the input image $I$ using the result $G_t$ of the previous iteration step as the guidance image:

$$G_{i+1} = \frac{1}{K_i} \sum_{j \in \Omega} I_j \cdot f(||i-j||) \cdot g(||G_i - G_j||).$$  \hspace{1cm} (11)

In this scheme, details smaller than the Gaussian kernel of the bilateral filter are removed while the edges of the remaining details are iteratively restored. Hence, this scheme allows the selective elimination of small scale details while preserving the remaining image structure. Note that the initial guidance image $G^1$ can simply be a constant (e.g., zero) valued image since it updates to the Gaussian filtered input image in the first iteration step. Here we propose to replace the bilateral filter in this scheme by a guided filter to avoid any gradient reversal artefacts.
RELATED WORK

As mentioned before, most multi-scale transform-based image fusion methods introduce some artefacts because the spatial consistency is not well-preserved (Li, Kang & Hu, 2013). This has led to the use of edge preserving filters to decompose source images into approximate and residual layers while preserving the edge information in the fusion process. Techniques that have been applied include weighted least squares filter (Yong & Minghui, 2014), $L_1$ fidelity using $L_0$ gradient (Cui et al., 2015), $L_0$ gradient minimization (Zhao et al., 2013), cross bilateral filter (Kumar, 2013) and anisotropic diffusion (Bavirisetti & Dhuli, 2016a).

Li, Kang & Hu (2013) proposed to restore spatial consistency by using guided filtering in the weighted recombination stage of the fusion process. In their scheme, the input images are first decomposed into approximate and residual layers using a simple averaging filter. Next, each input image is then filtered with a Laplacian kernel followed by blurring with a Gaussian kernel, and the absolute value of the result is adopted as a saliency map that characterizes the local distinctness of the input image details. Then, binary weight maps are obtained by comparing the saliency maps of all input images, and assigning a pixel in an individual weight map the value 1 if it is the pixelwise maximum of all saliency maps, and 0 otherwise. The resulting binary weight maps are typically noisy and not aligned with object boundaries and may produce artefacts to the fused image. Li, Kang & Hu (2013) performed guided filtering on each weight map with its corresponding source layer as the guidance image, to reduce noise and to restore spatial consistency. The GF guarantees that pixels with similar intensity values have similar weights and weighting is not performed across edges. Typically a large filter size and a large blur degree are used to fuse the approximation layers, while a small filter size and a small blur degree are used to combine the residual layers. Finally, the fused image is obtained by weighted recombination of the individual source residual layers. Despite the fact that this method is efficient and can achieve state-of-the-art performance in most cases, it does not use edge preserving filtering in the decomposition stage and applies a saliency map that does not relate well to human visual saliency (Gan et al., 2015).

In their multi-scale image fusion framework Gan et al. (2015) apply edge preserving filtering in the decomposition stage to extract well-defined image details (i.e., to preserve their edges) and use guided filtering in the weighted recombination stage to reduce spatial inconsistencies introduced by the weighting maps used in the reconstruction stage (i.e., to prevent edge artefacts like halos). First, a nonlinear weighted least squares edge-preserving filter (Farbman et al., 2008) is used to decompose the source images into approximate and residual layers. Next, phase congruency is used to calculate saliency maps that characterize the local distinctness of the source image details. The rest of their scheme is similar to that of Li, Kang & Hu (2013): binary weight maps are obtained from pixelwise comparison of the saliency maps corresponding to the individual source images; guided filtering is applied to these binary weight maps to recue noise and restore spatial consistency, and the fused image is obtained by weighted recombination of the individual source residual layers.
PROPOSED METHOD

A flow chart of the proposed multi-scale decomposition fusion scheme is shown in Fig. 1. The algorithm consists of the following steps:

1. Iterative guided filtering is applied to decompose the source images into approximate layers (representing large scale variations) and residual layers (containing small scale variations).
2. Frequency-tuned filtering (Achanta et al., 2009) is used to generate saliency maps for the source images.
3. Binary weighting maps are computed as the pixelwise maximum of the individual source saliency maps.
4. Guided filtering is applied to each binary weighting map with its corresponding source as the guidance image to reduce noise and to restore spatial consistency.
5. The fused image is computed as a weighted recombination of the individual source residual layers.

In a hierarchical framework steps 1–4 are performed at multiple spatial scales. In this paper we used a 4 level decomposition obtained by filtering at three different spatial scales (see Fig. 1).

Figure 2 shows the intensified visual (II) and thermal infrared (IR) or near infrared (NIR) images together with the results of the proposed image fusion scheme, for the 12 different scenes that were used in the present study. We will now discuss the proposed fusion scheme in more detail.

Consider two co-registered source images $X_0(x,y)$ and $Y_0(x,y)$. The proposed scheme then applies iterative guided filtering (IGF) to the input images $X_i$ and $Y_i$ to obtain progressively coarser image representations $X_{i+1}$ and $Y_{i+1}$ ($i > 0$):

$$\text{IGF}(X_i, r_i, \varepsilon_i) = X_{i+1}; \quad i \in \{0, 1, 2\}$$

(12)

where the parameters $\varepsilon_i$ and $r_i$ represent respectively the range and the spatial variances of the guided filter at iteration step $i$. In this study the number of iteration steps is set to 4. By letting each finer scale image serve as the approximate layer for the preceding coarser scale image the successive size-selective residual layers $dX_i$ are simply obtained by subtraction as follows:

$$dX_i = X_i - X_{i+1}; \quad i \in \{0, 1, 2\}.$$  

(13)

Figure 3 shows the approximate and residual layers that are obtained this way for the tank scene (nr 10 in Fig. 2). The edge-preserving properties of the iterative guided filter guarantee a graceful decomposition of the source images into details at different spatial scales. The filter size and regularization parameters used in this study are respectively set to $r_i = \{5, 10, 30\}$ and $\varepsilon_i = \{0.0001, 0.01, 0.1\}$ for $i = \{0, 1, 2\}$.

Visual saliency refers to the physical, bottom-up distinctness of image details (Fecteau & Munoz, 2006). It is a relative property that depends on the degree to which a detail is visually distinct from its background (Wertheim, 2010). Since saliency quantifies the relative visual importance of image details saliency maps are frequently used in the weighted recombination phase of multi-scale image fusion schemes (Bavirisetti & Dhuli, 2016b; Cui et al., 2015; Gan et al., 2015). Frequency tuned filtering computes bottom-up saliency as local multi-scale luminance contrast (Achanta et al., 2009). The saliency map $S$ for an image $I$ is computed as

$$S(x, y) = \| I_\mu - I_f (x, y) \|$$

(14)

where $I_\mu$ is the arithmetic mean image feature vector, $I_f$ represents a Gaussian blurred version of the original image, using a $5 \times 5$ separable binomial kernel, $\|\|$ is the $L_2$ norm (Euclidian distance), and $x, y$ are the pixel coordinates.
Figure 2  **Original input and fused images for all 12 scenes.** The intensified visual (II), thermal infrared (IR) or near infrared (NIR: scene 12) source images together with the result of the proposed fusion scheme (F) for each of the 12 scenes used in this study.

A recent and extensive evaluation study comparing 13 state-of-the-art saliency models found that the output of this simple saliency model correlates more strongly with human visual perception than the output produced by any of the other available models (*Toet, 2011*).
In the proposed fusion scheme we first compute saliency maps $S_{X_i}$ and $S_{Y_i}$ for the individual source layers $X_i$ and $Y_i$, $i \in \{0, 1, 2\}$. Binary weight maps $BW_{X_i}$ and $BW_{Y_i}$ are then computed by taking the pixelwise maximum of corresponding saliency maps $S_{X_i}$ and $S_{Y_i}$:

$$BW_{X_i}(x, y) = \begin{cases} 
1 & \text{if } S_{X_i}(x, y) > S_{Y_i}(x, y) \\
0 & \text{otherwise}
\end{cases}$$

$$BW_{Y_i}(x, y) = \begin{cases} 
1 & \text{if } S_{Y_i}(x, y) > S_{X_i}(x, y) \\
0 & \text{otherwise}.
\end{cases}$$

The resulting binary weight maps are noisy and typically not well aligned with object boundaries, which may give rise to artefacts in the final fused image. Spatial consistency is therefore restored through guided filtering (GF) of these binary weight maps with the corresponding source layers as guidance images:

$$W_{X_i} = GF(BW_{X_i}, X_i)$$
$$W_{Y_i} = GF(BW_{Y_i}, Y_i).$$

As noted before guided filtering combines noise reduction with edge preservation, while the output is locally approximately a scaled version of the guidance image. In the present scheme these properties are used to transform the binary weight maps into smooth
continuous weight maps through guided filtering with the corresponding source images as guidance images. Figure 4 illustrates the process of computing smoothed weight maps by guided filtering of the binary weight maps resulting from the pointwise maximum of the corresponding source layer saliency maps for the tank scene.

Fused residual layers are then computed as the normalized weighted mean of the corresponding source residual layers:

$$d_{Fi} = \frac{W_{Xi} \cdot dX_i + W_{Yi} \cdot dY_i}{W_{Xi} + W_{Yi}}.$$  \hspace{1cm} (17)

The fused image $F$ is finally obtained by adding the fused residual layers to the average value of the coarsest source layers:

$$F = \frac{X_3 + Y_3}{2} + \sum_{i=0}^{2} d_{Fi}.$$  \hspace{1cm} (18)

By using guided filtering both in the decomposition stage and in the recombination stage, this proposed fusion scheme optimally benefits from both the multi-scale edge-preserving characteristics (in the iterative framework) and the structure restoring capabilities (through guidance by the original source images) of the guided filter. The method is easy to implement and computationally efficient.

**METHODS AND MATERIAL**

This section presents the test imagery and computational metrics used to assess the performance of the proposed images fusion scheme in comparison to existing multi-scale fusion schemes.
Figure 5  **Comparison with existing multiresolution fusion schemes.** Original intensified visual (A) and thermal infrared (B) images for scene nr 10, and the fused results obtained with respectively a Contrast Pyramid (C), Gradient Pyramid (D), Laplace Pyramid (E), Morphological Pyramid (F), Ratio Pyramid (G), DWT (H), SIDWT (I), and the proposed method (J), for scene nr. 10.

**Test imagery**

Figure 2 shows the intensified visual (II), thermal infrared (IR) or near infrared (NIR: scene 12) source images together with the result of the proposed fusion scheme (F) for each of the 12 scenes used in this study. The 12 scenes are part of the TNO Image Fusion Dataset (*Toet, 2014*) with the following identifiers: airplane_in_trees, Barbed_wire_2, Jeep, Kaptein_1123, Marne_07, Marne_11, Marne_15, Reek, tank, Nato_camp_sequence, soldier_behind_smoke, Vlasakkers.

**Multi-scale fusion schemes used for comparison**

In this study we compare the performance of our image fusion scheme with seven other popular image fusion methods based on multi-scale decomposition including the Laplacian pyramid (*Burt & Adelson, 1983*), the Ratio of Low-Pass pyramid (*Toet, 1989b*), the contrast pyramid (*Toet, Van Ruyven & Valeton, 1989*), the filter-subtract-decimate Laplacian pyramid (*Burt, 1988; Burt & Kolczynski, 1993*), the gradient pyramid (*Burt, 1992; Burt & Kolczynski, 1993*), the morphological pyramid (*Toet, 1989a*), the discrete wavelet transform (*Lemeshewsky, 1999; Li, Manjunath & Mitra, 1995; Li, Kwok & Wang, 2002; Scheunders & De Backer, 2001*), and a shift invariant extension of the discrete wavelet transform (*Lemeshewsky, 1999; Rockinger, 1997; Rockinger, 1999; Rockinger & Fechner, 1998*). We used Rockinger’s freely available Matlab image fusion toolbox ([www.metapix.de/toolbox.htm](http://www.metapix.de/toolbox.htm)) to compute these fusion schemes. To allow a straightforward comparison, the number of scale levels is set to 4 in all methods, and simple averaging is used to compute the approximation of the fused image representation at the coarsest spatial scale. Figures 5–9 show the results of the proposed method together with the results of other seven fusion schemes for some of the scenes used in this study (scenes 2–5 and 10).

**Objective evaluation metrics**

Image fusion results can be evaluated using either subjective or objective measures. Subjective methods are based on psycho-visual testing and are typically expensive in terms
Figure 6  As Fig. 5, for scene nr. 2.

Figure 7  As Fig. 5, for scene nr. 3.

Figure 8  As Fig. 5, for scene nr. 4.
of time, effort, and equipment required. Also, in most cases, there is only little difference among fusion results. This makes it difficult to subjectively perform the evaluation of fusion results. Therefore, many objective evaluation methods have been developed (for an overview see e.g., Li, Li & Gong, 2010; Liu et al., 2012). However, so far, there is no universally accepted metric to objectively evaluate the image fusion results. In this paper, we use four frequently applied computational metrics to objectively evaluate and compare the performance of different image fusion methods. The metrics we use are Entropy, the Mean Structural Similarity Index (MSSIM), Normalized Mutual Information (NMI), and Normalized Feature Mutual Information (NFMI). These metrics will be briefly discussed in the following sections.

**Entropy**
Entropy ($E$) is a measure of the information content in a fused image $F$. Entropy is defined as

$$E_F = -\sum_{i=0}^{L-1} P_F(i) \log P_F(i)$$

where $P_F(i)$ indicates the probability that a pixel in the fused image $F$ has a gray value $i$, and the gray values range from 0 to $L$. The larger the entropy is, the more informative the fused image is. A fused image is more informative than either of its source images when its entropy is higher than the entropy of its source images.

**Mean Structural Similarity Index**
The Structural Similarity (SSIM: Wang et al., 2004) index is a stabilized version of the Universal Image Quality Index (UIQ: Wang & Bovik, 2002) which can be used to quantify the structural similarity between a source image $A$ and a fused image $F$:

$$\text{SSIM}_{x,y} = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \cdot \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \cdot \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$$

![Figure 9](image) As Fig. 5, for scene nr. 5.
where \( x \) and \( y \) represent local windows of size \( M \times N \) in respectively \( A \) and \( F \), and

\[
\mu_x = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} x(i,j), \quad \mu_y = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} y(i,j)
\]

\[
\sigma_x^2 = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} (x(i,j) - \mu_x)^2, \quad \sigma_y^2 = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} (y(i,j) - \mu_y)^2
\]

\[
\sigma_{xy}^2 = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} (x(i,j) - \mu_x)(y(i,j) - \mu_y).
\]

By default, the stabilizing constants are set to \( C_1 = (0.01 \cdot L)^2 \), \( C_2 = (0.03 \cdot L)^2 \) and \( C_3 = C_2/2 \), where \( L \) is the maximal gray value. The value of SSIM is bounded and ranges between \(-1\) and \( 1 \) (it is 1 only when both images are identical). The SSIM is typically computed over a sliding window to compare local patterns of pixel intensities that have been normalized for luminance and contrast. The Mean Structural Similarity (MSSIM) index quantifies the overall similarity between a source image \( A \) and a fused image \( F \):

\[
\text{MSSIM}_{A,F} = \frac{1}{N_w} \sum_{i=1}^{N_w} \text{SSIM}_{x_i,y_i}
\]

where \( N_w \) represents the number of local windows of the image. An overall image fusion quality index can then be defined as the mean MSSIM values between each of the source images and the fused result:

\[
\text{MSSIM}_{F}^{A,B} = \frac{\text{MSSIM}_{A,F} + \text{MSSIM}_{B,F}}{2}
\]

\( \text{MSSIM}_{F}^{A,B} \) ranges between \(-1\) and \( 1 \) (it is 1 only when both images are identical).

**Normalized Mutual Information**

Mutual Information (MI) measures the amount of information that two images have in common. It can be used to quantify the amount of information from a source image that is transferred to a fused image (Qu, Zhang & Yan, 2002). The mutual information \( \text{MI}_{AF} \) between a source image \( A \) and a fused image \( F \) is defined as:

\[
\text{MI}_{A,F} = \sum_{i,j} P_{A,F}(i,j) \log \frac{P_{A,F}(i,j)}{P_A(i)P_F(j)}
\]

where \( P_A(i) \) and \( P_F(j) \) are the probability density functions in the individual images, and \( P_{A,F}(i,j) \) is the joint probability density function.

The traditional mutual information metric is unstable and may bias the measure towards the source image with the highest entropy. This problem can be resolved by computing the normalized mutual information (NMI) as follows (Hossny, Nahavandi & Creighton, 2008):

\[
\text{NMI}_{F}^{A,B} = \frac{\text{MI}_{A,F}}{H_A + H_F} + \frac{\text{MI}_{B,F}}{H_B + H_F}
\]
where $H_A, H_B$ and $H_F$ are the marginal entropy of $A$, $B$ and $F$, and $\text{MI}_{A,F}$ and $\text{MI}_{B,F}$ represent the mutual information between respectively the source image $A$ and the fused image $F$ and between the source image $B$ and the fused image $F$. A higher value of NMI indicates that more information from the source images is transferred to the fused image. The NMI metric varies between 0 and 1.

**Normalized Feature Mutual Information**

The Feature Mutual Information (FMI) metric calculates the amount of image features that two images have in common (Haghighat & Razian, 2014; Haghighat, Aghagolzadeh & Seyedarabi, 2011). This method outperforms other metrics (e.g., E, NMI) in consistency with the subjective quality measures. Previously proposed MI-based image fusion quality metrics use the image histograms to compute the amount of information a source and fused image have in common (Cvejic, Canagarajah & Bull, 2006; Qu, Zhang & Yan, 2002). However, image histograms contain no information about local image structure (spatial features or local image quality) and only provide statistical measures of the number of pixels in a specific gray-level. However, since meaningful image information is contained in visual features, image fusion quality measures should measure the extent to which these visual features are transferred into the fused image from each of the source images. The Feature Mutual Information (FMI) metric calculates the mutual information between image feature maps (Haghighat & Razian, 2014; Haghighat, Aghagolzadeh & Seyedarabi, 2011). A typical image feature map is for instance the gradient map, which contains information about the pixel neighborhoods, edge strength and directions, texture and contrast. Given two source images as $A$ and $B$ and their fused image as $F$, the FMI metric first extracts feature maps of the source and fused images using a feature extraction method (e.g., gradient). After feature extraction, the feature images $A'$, $B'$ and $F'$ are normalized to create their marginal probability density functions $P_{A'}$, $P_{B'}$ and $P_{F'}$. The joint probability density functions $P_{A',F'}$ and $P_{B',F'}$ are then estimated from the marginal distributions using Nelsen’s method (Nelsen, 1987). The algorithm is described in more detail elsewhere (Haghighat, Aghagolzadeh & Seyedarabi, 2011). The FMI metric between a source image $A$ and a fused image $F$ is then given by

$$\text{FMI}_{A,F} = \text{MI}_{A',F'} = \sum_{i,j} P_{A',F'}(i,j) \log \frac{P_{A',F'}(i,j)}{P_{A'}(i)P_{F'}(j)}$$ \hspace{0.5cm} (28)

and the normalized feature mutual information (FMI) can be computed as follows

$$\text{FMI}_{F}^{A,B} = \frac{\text{MI}_{A',F'}}{H_A + H_F} + \frac{\text{MI}_{B',F'}}{H_B + H_F}.$$ \hspace{0.5cm} (29)

In practice the FMI is computed locally over small corresponding windows between the source and the fused images and averaged over all windows covering the image plane (Haghighat & Razian, 2014).
Table 1 | Entropy values for each of the methods tested and for all 12 scenes.

<table>
<thead>
<tr>
<th>Scene nr.</th>
<th>Contrast</th>
<th>DWT</th>
<th>Gradient</th>
<th>Laplace</th>
<th>Morph</th>
<th>Ratio</th>
<th>SIDWT</th>
<th>IGF</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>7.3335</td>
<td>7.3744</td>
<td>7.3379</td>
<td>7.3907</td>
<td><strong>7.4251</strong></td>
<td>7.3486</td>
<td>7.3746</td>
<td>7.3568</td>
</tr>
</tbody>
</table>

RESULTS

Fusion evaluation

Here we assess the performance of the proposed image fusion scheme on the intensified visual and thermal infrared images for each of the 12 selected scenes, using Entropy, the Mean Structural Similarity Index (MSSIM), Normalized Mutual Information (NMI), and Normalized Feature Mutual Information (NFMI) as the objective performance measures. We also compare the results of the proposed method with those of seven other popular multi-scale fusion schemes.

Table 1 lists the entropy of the fused result for the proposed method (IGF) and all seven multi-scale comparison methods (Contrast Pyramid, DWT, Gradient Pyramid, Laplace Pyramid, Morphological Pyramid, Ratio Pyramid, SIDWT). It appears that IGF produces a fused image with the highest entropy for 9 of the 12 test scenes. Note that a larger entropy implies more edge information, but it does not mean that the additional edges are indeed meaningful (they may result from over enhancement or noise). Therefore, we also need to consider structural information metrics.

Table 2 shows that IGF outperforms all other multi-scale methods tested here in terms of MSSIM. This means that the mean overall structural similarity between both source images the fused image $F$ is largest for the proposed method.

Table 3 shows that IGF also outperforms all other multi-scale methods tested here in terms of NMI. This indicates that the proposed IGF fusion scheme transfers more information from the source images to the fused image than any of the other methods.

Table 4 shows that IGF also outperforms 10 of the 12 other multi-scale methods tested here in terms of NFMI. IGF is only outperformed by SIDWT for scene 1 and by the Contrast Pyramid for scene 7. This implies that fused images produced by the proposed IGF scheme typically have a larger amount of image features in common with their source images than the results of most other fusion schemes.
Table 2  MSSIM values for each of the methods tested and for all 12 scenes.

<table>
<thead>
<tr>
<th>Scene nr.</th>
<th>Contrast</th>
<th>DWT</th>
<th>Gradient</th>
<th>Laplace</th>
<th>Morph</th>
<th>Ratio</th>
<th>SIDWT</th>
<th>IGF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7851</td>
<td>0.7975</td>
<td>0.8326</td>
<td>0.8050</td>
<td>0.7321</td>
<td>0.8054</td>
<td>0.8114</td>
<td>0.8381</td>
</tr>
<tr>
<td>2</td>
<td>0.6018</td>
<td>0.6798</td>
<td>0.7130</td>
<td>0.6406</td>
<td>0.6203</td>
<td>0.6406</td>
<td>0.6935</td>
<td>0.7213</td>
</tr>
<tr>
<td>3</td>
<td>0.7206</td>
<td>0.7493</td>
<td>0.7849</td>
<td>0.7555</td>
<td>0.6882</td>
<td>0.7468</td>
<td>0.7629</td>
<td>0.7932</td>
</tr>
<tr>
<td>4</td>
<td>0.6401</td>
<td>0.6790</td>
<td>0.7162</td>
<td>0.6875</td>
<td>0.6155</td>
<td>0.6668</td>
<td>0.6949</td>
<td>0.7184</td>
</tr>
<tr>
<td>5</td>
<td>0.5856</td>
<td>0.6649</td>
<td>0.6938</td>
<td>0.6695</td>
<td>0.6250</td>
<td>0.6270</td>
<td>0.6769</td>
<td>0.7038</td>
</tr>
<tr>
<td>6</td>
<td>0.5689</td>
<td>0.6448</td>
<td>0.6755</td>
<td>0.6516</td>
<td>0.5961</td>
<td>0.6099</td>
<td>0.6598</td>
<td>0.6921</td>
</tr>
<tr>
<td>7</td>
<td>0.3939</td>
<td>0.5742</td>
<td>0.5994</td>
<td>0.5809</td>
<td>0.5320</td>
<td>0.4490</td>
<td>0.5889</td>
<td>0.6344</td>
</tr>
<tr>
<td>8</td>
<td>0.6474</td>
<td>0.6272</td>
<td>0.6630</td>
<td>0.6392</td>
<td>0.5791</td>
<td>0.6291</td>
<td>0.6463</td>
<td>0.6940</td>
</tr>
<tr>
<td>9</td>
<td>0.6224</td>
<td>0.6883</td>
<td>0.7224</td>
<td>0.6955</td>
<td>0.6445</td>
<td>0.6718</td>
<td>0.7089</td>
<td>0.7405</td>
</tr>
<tr>
<td>10</td>
<td>0.3913</td>
<td>0.5410</td>
<td>0.5715</td>
<td>0.5430</td>
<td>0.4899</td>
<td>0.4331</td>
<td>0.5513</td>
<td>0.5961</td>
</tr>
<tr>
<td>11</td>
<td>0.7174</td>
<td>0.7307</td>
<td>0.7754</td>
<td>0.7439</td>
<td>0.6559</td>
<td>0.7419</td>
<td>0.7539</td>
<td>0.7908</td>
</tr>
<tr>
<td>12</td>
<td>0.7945</td>
<td>0.8116</td>
<td>0.8466</td>
<td>0.8227</td>
<td>0.7815</td>
<td>0.8106</td>
<td>0.8365</td>
<td>0.8646</td>
</tr>
</tbody>
</table>

Table 3  NMI values for each of the methods tested and for all 12 scenes.

<table>
<thead>
<tr>
<th>Scene nr.</th>
<th>Contrast</th>
<th>DWT</th>
<th>Gradient</th>
<th>Laplace</th>
<th>Morph</th>
<th>Ratio</th>
<th>SIDWT</th>
<th>IGF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1534</td>
<td>0.1692</td>
<td>0.2052</td>
<td>0.1647</td>
<td>0.1699</td>
<td>0.1791</td>
<td>0.1796</td>
<td>0.2818</td>
</tr>
<tr>
<td>2</td>
<td>0.0989</td>
<td>0.0948</td>
<td>0.1158</td>
<td>0.0897</td>
<td>0.1028</td>
<td>0.0897</td>
<td>0.1028</td>
<td>0.2994</td>
</tr>
<tr>
<td>3</td>
<td>0.0898</td>
<td>0.1222</td>
<td>0.1493</td>
<td>0.1252</td>
<td>0.1171</td>
<td>0.1320</td>
<td>0.1280</td>
<td>0.2231</td>
</tr>
<tr>
<td>4</td>
<td>0.1102</td>
<td>0.1097</td>
<td>0.1322</td>
<td>0.1189</td>
<td>0.1169</td>
<td>0.1046</td>
<td>0.1177</td>
<td>0.2294</td>
</tr>
<tr>
<td>5</td>
<td>0.1236</td>
<td>0.1170</td>
<td>0.1379</td>
<td>0.1252</td>
<td>0.1318</td>
<td>0.1186</td>
<td>0.1251</td>
<td>0.2166</td>
</tr>
<tr>
<td>6</td>
<td>0.0857</td>
<td>0.0943</td>
<td>0.1162</td>
<td>0.0969</td>
<td>0.1068</td>
<td>0.0902</td>
<td>0.0980</td>
<td>0.2229</td>
</tr>
<tr>
<td>7</td>
<td>0.0697</td>
<td>0.0711</td>
<td>0.0839</td>
<td>0.0809</td>
<td>0.0888</td>
<td>0.0616</td>
<td>0.0781</td>
<td>0.2147</td>
</tr>
<tr>
<td>8</td>
<td>0.2192</td>
<td>0.1825</td>
<td>0.2198</td>
<td>0.1832</td>
<td>0.1884</td>
<td>0.2130</td>
<td>0.2021</td>
<td>0.3090</td>
</tr>
<tr>
<td>9</td>
<td>0.0692</td>
<td>0.0679</td>
<td>0.0781</td>
<td>0.0747</td>
<td>0.0790</td>
<td>0.0690</td>
<td>0.0731</td>
<td>0.2013</td>
</tr>
<tr>
<td>10</td>
<td>0.1375</td>
<td>0.1643</td>
<td>0.2043</td>
<td>0.1780</td>
<td>0.1761</td>
<td>0.1662</td>
<td>0.1760</td>
<td>0.2962</td>
</tr>
<tr>
<td>11</td>
<td>0.1055</td>
<td>0.1043</td>
<td>0.1177</td>
<td>0.1100</td>
<td>0.1047</td>
<td>0.1179</td>
<td>0.1115</td>
<td>0.1646</td>
</tr>
<tr>
<td>12</td>
<td>0.2572</td>
<td>0.2511</td>
<td>0.2746</td>
<td>0.2602</td>
<td>0.2438</td>
<td>0.2660</td>
<td>0.2649</td>
<td>0.2987</td>
</tr>
</tbody>
</table>

Summarizing, the proposed IGF fusion scheme appears to outperform the other multi-scale fusion methods investigated here in most of the conditions tested.

**Runtime**

In this study we used a Matlab implementation of the GF and IGF written by Zhang *et al.* (2014) that is freely available from the authors (at http://www.cs.cuhk.edu.hk/~leojia/projects/rollguidance). We made no effort to optimize the code of the algorithms. We conducted a runtime test on a Dell Latitude laptop with an Intel i5 2 GHz CPU and 8 GB memory. The algorithms were implemented in Matlab 2016a. Only a single thread was used without involving any SIMD instructions. For this test we used the set of 12 test images described in ‘Test imagery.’ As noted before, the filter size and regularization parameters used in this study are respectively set to $r_i = \{5, 10, 30\}$ and $\epsilon_i = \{0.0001, 0.01, 0.1\}$ for
Table 4  NFMI values for each of the methods tested and for all 12 scenes.

<table>
<thead>
<tr>
<th>Scene nr.</th>
<th>Contrast</th>
<th>DWT</th>
<th>Gradient</th>
<th>Laplace</th>
<th>Morph</th>
<th>Ratio</th>
<th>SIDWT</th>
<th>IGF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.4064</td>
<td>0.3812</td>
<td>0.3933</td>
<td>0.3888</td>
<td>0.3252</td>
<td>0.3498</td>
<td><strong>0.4084</strong></td>
<td>0.4008</td>
</tr>
<tr>
<td>2</td>
<td>0.4354</td>
<td>0.3876</td>
<td>0.4001</td>
<td>0.3493</td>
<td>0.3432</td>
<td>0.3493</td>
<td>0.4075</td>
<td><strong>0.4383</strong></td>
</tr>
<tr>
<td>3</td>
<td>0.4076</td>
<td>0.4081</td>
<td>0.4175</td>
<td>0.4138</td>
<td>0.3758</td>
<td>0.3552</td>
<td>0.4330</td>
<td><strong>0.4454</strong></td>
</tr>
<tr>
<td>4</td>
<td>0.4017</td>
<td>0.3913</td>
<td>0.4066</td>
<td>0.4051</td>
<td>0.3655</td>
<td>0.3497</td>
<td>0.4205</td>
<td><strong>0.4490</strong></td>
</tr>
<tr>
<td>5</td>
<td>0.4304</td>
<td>0.3971</td>
<td>0.4101</td>
<td>0.4081</td>
<td>0.3758</td>
<td>0.3497</td>
<td>0.4229</td>
<td><strong>0.4580</strong></td>
</tr>
<tr>
<td>6</td>
<td>0.4299</td>
<td>0.4074</td>
<td>0.4203</td>
<td>0.4164</td>
<td>0.3832</td>
<td>0.3570</td>
<td>0.4295</td>
<td><strong>0.4609</strong></td>
</tr>
<tr>
<td>7</td>
<td><strong>0.5050</strong></td>
<td>0.4383</td>
<td>0.4439</td>
<td>0.4357</td>
<td>0.3942</td>
<td>0.3779</td>
<td>0.4469</td>
<td>0.4286</td>
</tr>
<tr>
<td>8</td>
<td>0.4305</td>
<td>0.4074</td>
<td>0.4097</td>
<td>0.4113</td>
<td>0.3806</td>
<td>0.3553</td>
<td>0.4273</td>
<td><strong>0.4325</strong></td>
</tr>
<tr>
<td>9</td>
<td>0.4351</td>
<td>0.3959</td>
<td>0.4105</td>
<td>0.3995</td>
<td>0.3658</td>
<td>0.3539</td>
<td>0.4130</td>
<td><strong>0.4370</strong></td>
</tr>
<tr>
<td>10</td>
<td>0.4439</td>
<td>0.4251</td>
<td>0.4263</td>
<td>0.4268</td>
<td>0.3863</td>
<td>0.3465</td>
<td>0.4513</td>
<td><strong>0.5045</strong></td>
</tr>
<tr>
<td>11</td>
<td>0.3882</td>
<td>0.3798</td>
<td>0.3987</td>
<td>0.3804</td>
<td>0.3131</td>
<td>0.3453</td>
<td>0.4068</td>
<td><strong>0.4206</strong></td>
</tr>
<tr>
<td>12</td>
<td>0.4051</td>
<td>0.3725</td>
<td>0.3973</td>
<td>0.3820</td>
<td>0.3449</td>
<td>0.3635</td>
<td>0.4111</td>
<td><strong>0.4257</strong></td>
</tr>
</tbody>
</table>

spatial scale levels \( i = \{0, 1, 2\} \). The mean runtime of the proposed fusion method was 0.61 ± 0.05 s.

DISCUSSION AND CONCLUSIONS

We propose a multi-scale image fusion scheme based on guided filtering. Iterative guided filtering is used to decompose the source images into approximation and residual layers. Initial binary weighting maps are computed as the pixelwise maximum of the individual source saliency maps, obtained from frequency tuned filtering. Spatially consistent and smooth weighting maps are then obtained through guided filtering of the binary weighting maps with their corresponding source layers as guidance images. Saliency weighted recombination of the individual source residual layers and the mean of the coarsest scale source layers finally yields the fused image. The proposed multi-scale image fusion scheme achieves spatial consistency by using guided filtering both at the decomposition and at the recombination stage of the multi-scale fusion process. Application to multiband visual (intensified) and thermal infrared imagery demonstrates that the proposed method obtains state-of-the-art performance for the fusion of multispectral nightvision images. The method has a simple implementation and is computationally efficient.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding
The effort was sponsored by the Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant number FA9550-15-1-0433. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Grant Disclosures
The following grant information was disclosed by the author:

Competing Interests
The author declares there are no competing interests.

Author Contributions
• Alexander Toet conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, performed the computation work, reviewed drafts of the paper.

Data Availability
The following information was supplied regarding data availability:
Figshare: TNO Image Fusion Dataset
http://dx.doi.org/10.6084/m9.figshare.1008029.

REFERENCES


Xue Z, Blum RS, Li Y. 2002. Fusion of visual and IR images for concealed weapon
IEEE, 1198–1205.

conference on computational intelligence and software engineering 2009

IEEE international conference on medical imaging physics and engineering (ICMIPE).*


Yong J, Minghui W. 2014. Image fusion using multiscale edge-preserving decompo-
sition based on weighted least squares filter. *IET Image Processing* 8(3):183–190

Zhang B, Lu X, Pei H, Zhao Y. 2015. A fusion algorithm for infrared and visible images
based on saliency analysis and non-subsampled Shearlet transform. *Infrared Physics


visual weight map extraction based on multi scale edge preserving decomposition.


Zitová B, Beneš M, Blažek J. 2011. Image fusion for art analysis. In: *Computer vision and
Optical Engineering, 7869081–7869089.

international workshop on object tracking and classification in and beyond the visible
spectrum (OTCBVS’05).* Piscataway: IEEE, W01-30-01-08.