Infrared and Visible Image Fusion based on Saliency Detection and Infrared Target Segment

JUN LI, MINGHUI SONG and YUANXI PENG

ABSTRACT

Infrared and visible image fusion is significant to the overall detection and tracking performance of a video surveillance system. Considering the characteristics of infrared and visible images, the ideal fusion of the infrared and visual image should integrate the bright features of the infrared image, while preserving a considerable amount of background information of the visible image. However, current methods use inadequate information extraction. An efficient infrared and visible image fusion method is thus proposed in this paper based on saliency detection and infrared target segmentation. Firstly, an image descriptor, referred to as an image signature, is introduced. The saliency algorithm is implemented to obtain the infrared image region of interest. Secondly, an infrared target is segmented from the region of interest by adopting a thresholding method. Subsequently, the infrared target is added to the image, which is fused from the original infrared and visible image by known image fusion methods. The most remarkable advantage of the proposed method is that it can obtain a clear infrared target. To verify the proposed method, it was experimentally compared with several classic fusion algorithms. By subjectively comparing the fused results, it was demonstrated that the proposed method not only extracted the important infrared targets, but it also showed good visual quality retained in the background information.

KEYWORDS

International Conference, International Conference, International Conference.

INTRODUCTION

Image fusion is designed to combine multiple input images into a fused image, which is expected to be more informative for human and machine perception compared to any of the input images [1]. Image fusion technology has been applied in various fields. Infrared and visible image fusion play an especially significant role in military actions and civil surveillance [2]. In light and infrared imaging technology, infrared images can display thermal targets under conditions of low lighting and occlusion, such as by trees or smog. However, infrared images are blurry and thus ineffectively describe details. On the contrary, a visible image is usually suitable for the human visual system. It has refined sharpness and high spatial resolution, which can provide the geometry and texture detail information of the scene [3-6]. It is apparent that the
infrared image and visible image offer complementary advantages. Therefore, development of effective infrared and visual image fusion methods are needed.

Current image fusion mainly focuses on the multiscale transform (MST)[1]. Traditional MST-based fusion algorithms are pyramid-based, such as the Laplacian pyramid (LP)[7], wavelet-based, such as the discrete wavelet transform (DWT)[8], multiscale geometric analysis (MGA)-based, such as the curvelet transform (CVT)[9], and relatively new methods, such as sparse representation (SR)[10], non-subsampled contourlet transform (NSCT)[11], and non-subsampled shearlet transform (NSST)[12].

When these methods are applied to fuse infrared and visible images, their fusion results are usually appealing to perception by the human visual system to some extent. However, an obvious and common disadvantage is that the infrared target of the fused image has low contrast and suffers from a blurring effect. To solve this problem, Zhang et al.[13] proposed an infrared and visual image fusion algorithm using infrared feature extraction and visual information preservation. In this method, infrared features are simply added to the visible image, which is ineffective in certain conditions, such as smog. In addition, Bavirisetti et al.[14] proposed an image fusion method based on saliency detection and double scale image decomposition. However, their weight map construction process can hardly exclude some other salient regions, which have no relation to the infrared target. Furthermore, Zhang et al.[15] proposed a fusion algorithm for infrared and visible images based on saliency analysis and non-subsampled shearlet transform. That super-pixel-based saliency analysis method is used to extract the salient regions. Nevertheless, the fused image obtained by this method shows unclear infrared targets.

FUSION FRAMEWORK

The proposed method is comprised of three key procedures: region of interest detection, infrared target segmentation, and secondary fusion. The method is illustrated in Fig. 1. It can be summarized as follows:

- A saliency algorithm is implemented to obtain the region of interest of the infrared image. The region of interest is refined through a binary, holistic image descriptor called the image signature. It is defined as the sign function of the discrete cosine transform (DCT) of an image [14].

- The infrared target is segmented from the region of interest by adopting a thresholding method. Inspired by psychology research relating to visual attention, we firstly extract the infrared feature map of the region of interest. Then, we implement segmentation and clustering in the region of interest. Finally, we obtain the infrared target by adopting the thresholding method.

- The notion of secondary fusion is hence introduced. The infrared target obtained by saliency detection and infrared target segmentation is added to the image, which is fused from the original infrared and visible image by known image fusion methods.
PROPOSED ALGORITHM

Image Signature

It is observed in many infrared images, which are captured in low-light conditions, that the background is usually darker and blurrier than the bright target. Thus, the infrared image features can be extracted. The problem of finding objects in a scene and separating them from the background is known as figure–ground separation [16-19].

Suppose a grey-scale image described in the following structure:

\[ I = T + B, \quad I, T, B \in \mathbb{IR}^N \]  (1)

Were \( T \) represents the target in the foreground, which is usually assumed to be sparse in a spatial basis. \( B \) denotes the background, which is sparse in terms of the discrete cosine transform. This means that the values of \( T \) and \( \hat{B} = DCT(B) \) have only a small number of non-zero components.

To solve the figure–ground separation problem, we can approximately isolate the support of \( T \) by taking the sign of the mixture signal \( I \) in the transformed domain and then inversely transform it back into the spatial domain. “Image signature—\( Is \)” preferentially contains foreground information and is defined as:

\[ Is(X) = sign(DCT(X)) \]  (2)

The reconstructed image can be computed as:

\[ \hat{X} = IDCT[sign(\hat{X})] \]  (3)

The infrared target in the foreground is visually conspicuous in relation to the background. We can thus obtain saliency map \( m \) by smoothing the squared reconstructed image defined above[20].

\[ m = g \ast (\bar{X} \odot \hat{X}) \]  (4)

Figure 1. General framework of the proposed method.
Were \( g \) is a Gaussian kernel, the asterisk \( * \) represents the convolution operator, and the dot/circle \( \odot \) denotes the Hadamard (entrywise) product operator.

Figure 2 illustrates the saliency map obtained through the image signature. In addition, Fig. 3(a) shows the saliency map overlapped on the original image, and Fig. 3(b) is the most salient part of the original image. The region of interest is the neighborhood of the most salient part of the original image, which is shown in Fig. 4(a).

**Feature Extraction and Selection**

The texture map can be obtained by calculating the standard variance of a pixel’s \( N \times N \) neighborhood window. In the texture map, the regional boundary point has a larger grey value, which is considered a bright point. The point that is not in the regional boundary is considered dark. Therefore, the texture map can effectively reflect the local features of an image. We extract the bright point in the texture map as the feature point of the region of interest. Then, the feature points are selected and clustered to obtain the center of the region of interest [21].

**ESTABLISHING THE TEXTURE MAP**

The variance of a pixel \( (m,n) \) is defined as:

\[
V(m,n) = \frac{1}{N} \sqrt{\sum_{i=n-M}^{n+M} \sum_{j=m-M}^{m+M} x(i,j)^2 - \frac{1}{N^2} \left( \sum_{i=n-M}^{n+M} \sum_{j=m-M}^{m+M} x(i,j) \right)^2}
\]  

(5)

Where \( x(i,j) \) is the grey value of pixel \((i,j)\), and \((i,j)\) belongs to the \(N \times N\) sliding window; \(M = N/2\). By computing the variance of every sliding window, we can obtain the texture map, as shown in Fig. 4(b).
FEATURE POINTS OF THE REGION OF INTEREST

An adaptive algorithm based on image complexity is applied to extract the feature points of the region of interest. For an image size of $M \times N$, the average variance is:

$$VA = \frac{1}{M \times N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} V(i, j)$$

(6)

The average deviation of $VA$ is defined as $D$:

$$D = \sqrt{\frac{1}{M \times N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (V(i, j) - \overline{VA})^2}$$

(7)

Were $\overline{VA}$ denotes the average grey value of the overall image, and $\overline{D}$ represents the deviation between the texture map and $\overline{VA}$. For an image with a complex scene, there are more bright points in the texture map, which means a larger value of $\overline{VA}$ and $\overline{D}$. Therefore, we define threshold $T$ to extract the feature points, thereby extracting the brighter points with a stronger saliency [22]. Obviously, if there are too many feature points, the computation complexity increases and an incorrect feature may be obtained. On the other hand, if the number of feature points is not adequate, the region of interest may be missed. Thus, the threshold is very important. Here, we define $T = VA + D$, which can produce an excellent feature extraction effect.

$$C(i, j) = \begin{cases} V(i, j) & V(i, j) \geq T \\ 0 & V(i, j) < T \end{cases}$$

(8)

Were $C(i, j)$ denotes the grey values of points in the feature map, as shown in Fig. 4(c).

SELECTING AND CLUSTERING FEATURE POINTS

To obtain more accurate and effective feature points, selecting and clustering is necessary [23-24]. Here, we present an algorithm based on the directional gradient vector. It can be described in two steps, as shown in Fig. 5.
1. For each point in the feature map, obtain the directional gradient vector of its eight neighboring pixels.
2. If a pixel has six or more directional gradient vector points to itself, retain this feature point. Otherwise, the feature point should be abandoned.

Fig. 6 shows the selected optimal candidate feature points.

After selection, the adjacent feature points belong to the region of interest. Thus, we can obtain the approximate center by clustering based on the geometric distance.

Suppose \( C(x_i, y_i) (i = 1, 2, \ldots, m) \) comprises the selected feature points. The clustering algorithm can be described as follows. Make the first feature point \( C(x_1, y_1) \) the clustering center of category one, described as \( w_1 \). Clustering center \( z_1 = C(x_1, y_1) \). Compute the distance \( d_{z1} \) between the next feature point \( C(x_2, y_2) \) and clustering center of category one \( z_1 \). If \( d_{z1} > D \), set up a new category \( w_2 \) with a center of \( z_2 = C(x_2, y_2) \). Otherwise, determine that \( C(x_2, y_2) \) belongs to \( w_1 \), and then upgrade the geometry center of \( C(x_1, y_1) \) and \( C(x_2, y_2) \) as the clustering center of \( w_1 \). The remainder can be performed in the same manner until all of feature points are classified, and the final clustering center is the region of interest center, as shown in Fig. 7.

**Infrared Target Segmentation**

The threshold method is used to segment the infrared target from the region of interest. The algorithm of iteratively computing the threshold is comprised of five steps.

Step 1: Obtain the maximum grey value \( Z_{\text{max}} \) and minimum grey value \( Z_{\text{min}} \). Then, compute the initial threshold.

\[
T^0 = (Z_{\text{max}} + Z_{\text{min}}) / 2
\] (9)
Step 2: Divide the image into two parts according to threshold $T^k$, and separately compute the average grey value.

$$Z_A = \frac{\sum_{z(i,j) < T^k} z(i,j) \times N(i,j)}{\sum_{z(i,j) < T^k} N(i,j)} \quad Z_B = \frac{\sum_{z(i,j) > T^k} z(i,j) \times N(i,j)}{\sum_{z(i,j) > T^k} N(i,j)}$$

(10)

where $z(i,j)$ is the grey value of pixel $(i,j)$, and $N(i,j)$ is the weight coefficient.

Step 3: Upgrade the threshold: $T^{k+1} = \frac{Z_A + Z_B}{2}$.

Step 4: If $T^k = T^{k+1}$, end the iteration. Otherwise, $k = k + 1$ and return to Step 2.

Step 5: Obtain the optimal threshold. Perform the binarization segmentation as:

$$f_T(i,j) = \begin{cases} f(i,j), & f(i,j) \geq T \\ 0, & f(i,j) < T \end{cases}$$

(11)

Where $f(i,j)$ the grey is value of pixel $(i,j)$ in the original region of interest, and $f_T$ is the segmentation result, as shown in Fig. 8.

**EXPERIMENTAL AND ANALYSIS**

**Experimental Setup**

To verify the effectiveness of the proposed fusion framework, several groups of commonly used infrared and visible images were processed in an experiment. All test images were from a public image fusion database. For each pair, the two source images were assumed to be pre-registered. Seven typical image fusion algorithms were compared with the proposed framework: ratio of low-pass pyramid (RP), dual-tree complex wavelet transform (DTCWT), LP, DWT, CVT, NSCT, and NSST. Among these, LP, SR, DWT, DTCWT, CVT are classic fusion methods, while NSCT and NSST methods have become more widely used in recent years.

**Experimental Result**

Figure 9 show the fusion results on the first image set obtained by different image fusion algorithms. In this figure, (a) and (b) are the source visible and infrared images. In addition, (c), (e), (g), (i), (k), (m), and (o) are respectively the fused image based on LP, RP, DWT, DTCWT, CVT, NSCT, and NSST. In addition, (d), (f), (h), (j), (l), (n), and (p) are results of our framework using a different relevant algorithm. It is readily apparent that the results of the existing fusion algorithm have a shared disadvantage, which is poor discrimination between the infrared target and background. Our framework is adequate to solve this problem and provides a marked improvement of the visual effect.
To simply the paper, we only present the fusion result of another image set with NSST and “NSST-ours,” as shown in Figs. 10 to 13.

Figure 9. Comparison example on the first image set.

Figure 10. Comparison example on the second image set with the NSST algorithm.

Figure 11. Comparison example on the third image set with the NSST algorithm.

Figure 12. Comparison example on the fourth image set with the NSST algorithm.
CONCLUSION

Saliency detection reflects the visually significant part of a scene, which is generally where the target is located. This attribute precisely fits the need of target segmentation in infrared and visible image fusion. A method for infrared and visible images fusion based on saliency detection and infrared target segmentation was thus proposed in this paper. The proposed method integrates the advantages of both saliency detection and target segmentation. It solves the common problem of the typical image fusion algorithm, specifically the low contrast between the target and background. Experimental results demonstrated that the proposed framework can not only highlight the infrared image target, but it also provides abundant background information of the visible image. Moreover, results of extensive qualitative evaluations verified that the proposed algorithm outperformed the related state-of-the-art algorithms.

ACKNOWLEDGEMENTS

This work was financially supported by the Research Fund of State Key Laboratory of High Performance Computing, National University of Defense Technology, P.R. China, under grant No. 201612-01.

Corresponding author: Yuanxi Peng. College of Computer Science, National University of Defense Technology, Chang Sha, China

REFERENCES