Feature-based fusion of infrared and visible dynamic images using target detection

Congyi Liu (刘从义)[1], Zhongliang Jing (敬忠良)[2], Gang Xiao (肖 刚)[2], and Bo Yang (杨 波)[1]

[1]*School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200030*

[2]*Institute of Aerospace Science & Technology, Shanghai Jiao Tong University, Shanghai 200030*

Received September 26, 2006

We employ the target detection to improve the performance of the feature-based fusion of infrared and visible dynamic images, which forms a novel fusion scheme. First, the target detection is used to segment the source image sequences into target and background regions. Then, the dual-tree complex wavelet transform (DT-CWT) is proposed to decompose all the source image sequences. Different fusion rules are applied respectively in target and background regions to preserve the target information as much as possible. Real world infrared and visible image sequences are used to validate the performance of the proposed novel scheme. Compared with the previous fusion approaches of image sequences, the improvements of shift invariance, temporal stability and consistency, and computation cost are all ensured.

OCIS codes: 100.2000, 100.2980, 100.7410, 110.3080.

The techniques of multi-source image fusion originated in the military fields, and their impetuses also came from military fields. The battlefield detecting technology, based on the pivotal content of multi-sources image fusion, has become one of the most important military advanced technologies, including target detection, track, recognition, and scene awareness.

Image fusion must satisfy the following requirements[1]: preserve all relevant information (as much as possible) in the composite image; do not introduce any artifacts or inconsistencies; be shift and rotational invariant; be temporal stability and consistency. The later two requirements are especially important in dynamic images (or image sequences) fusion.

Image fusion can be performed at different levels of information representation, classified in ascending order of abstraction: signal, pixel, feature, and symbol levels[2]. Recently, static images pixel-based fusion methods have been researched extensively[3,4]. However, few researchers have recently done some work on the dynamic images (or image sequences) fusion. Even so, they just focused on the pixel-based fusion of the image sequences. In this paper, we will do the research on the feature-based fusion of the image sequences using the region target detection. It will get both qualitative and quantitative improvements compared with the pixel-level methods. Because it has more intelligent semantic fusion rules which can be considered based on actual features, it can preserve the target information as much as possible.

Figure 1 shows the generic pixel-based image fusion method, which can be divided into three steps as follows. First, all source images are decomposed by using multi-resolution (MR) method, which can be the discrete wavelet transform (DWT)[5], discrete wavelet frames (DWF)[1,2,6] etc.. Then, the decomposition coefficients are fused by applying some fusion rule, which can be a point-based maximum selection (MS) rule or more sophisticated area-based rules. Finally, the fused image is reconstructed by using the corresponding inverse transform.

For pixel-based approaches, the MR decomposition coefficients are treated independently (MS rule) or filtered by a small fixed window (area-based rule). However, the most applications of a fusion scheme are interested in features within the image, not the actual pixels. Therefore, it seems reasonable to incorporate feature information into the fusion process. Indeed, a number of feature-level fusion schemes have been proposed. However, most of them are designed for static image fusion, and every frame of each source sequence is processed individually in image sequences case. These methods do not take full advantage of the wealth of inter-frame-information within source sequences. We can make use of the advantage of the inconspicuous changes between the adjacent frames among the image sequences, and use the information of the former frame to supervise the process of recent frame. Therefore, not only can it increase the speed of the processing, but also make full use of the abundant inter-frame-information.

In this paper, we propose a novel scheme of feature-based fusion of infrared (IR) and visible image sequences using the target detection, as shown in Fig. 2, where the target detection (TD) technique is introduced to segment target regions intelligently. To be convenient, We assume

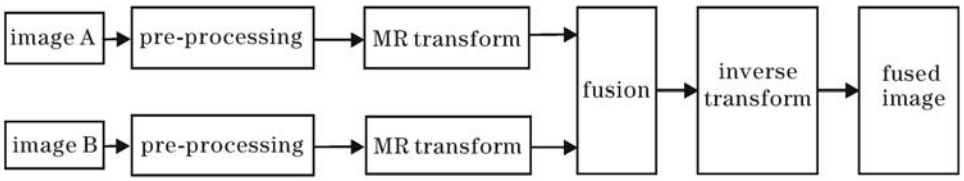

Fig. 1. Pixel-based image fusion scheme.

1671-7694/2007/050274-04

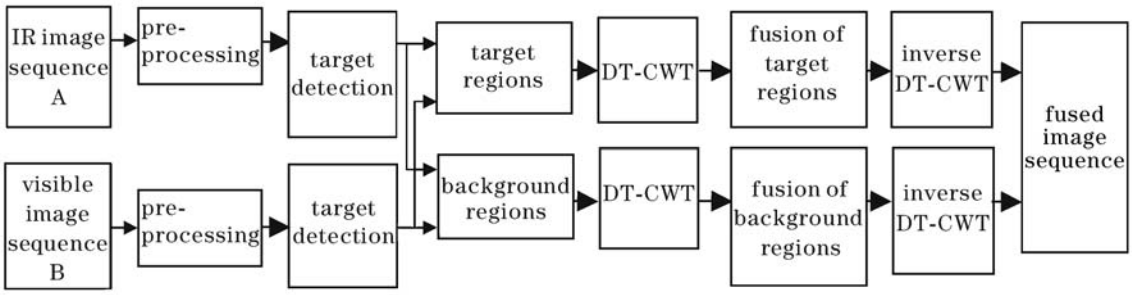

Fig. 2. Feature-based fusion of infrared and visible image sequences.

that both source sequences are well registered before their fusion. First, both of the visible and IR sequences are enhanced by using pre-processing operator. Secondly, each of the source images is segmented into target and background regions by using target detection. Then, we use the dual-tree complex wavelet transform (DT-CWT) to transform the source image sequences. Later, different fusion rules are respectively used in target and background regions. Finally, the fused coefficients belonging to each region are combined, and fused frames are reconstructed by using the corresponding inverse transform.

The TD operator aims for segmenting source frames into target and background regions. In the target regions, the significant information, such as moving human and vehicle, is included, and we need to preserve the target information as much as possible. In the later process of the fusion, we can respectively use different fusion rules between the target and background regions. In this case, we can preserve most of the information of the target regions which we are interested in.

At first, a fuzzy c-means clustering method[7] is adopted to segment the initial IR frame. In the segmented IR frame, it is easy to find the target regions, which have high contrast with the neighboring background. Then, a confidence measure[7] for each candidate region is computed. But it is very inefficient to compute the confidence measure for each candidate within every frame. Therefore, a model matching method is adopted to find the target regions in the subsequent frames. First, we record the center position of the target region in the pre-frame, and the target model is obtained by using intensity information of the target region in the pre-frame. Then, in the local area around the center position in the pre-frame, not the whole but a small region in the post-frame which corresponds with (and is little larger than) the target region in the pre-frame is matched to get the target region in the post-frame. Then the intensity information of the target region in the post-frame can be preserved as the target model. The initial detection operator based on segmentation and confidence measure will be repeated in case no target detected in some successive frames.

Here, the target detection is employed to use different fusion rules to preserve the target information as much as possible. The methods of detection, such as the fuzzy c-means clustering method, the confidence measure and model matching method, just need to compute in the gray level and are relatively simple, so they need low computation cost, which cannot notably increase the computational-complexity of the dynamic fusion scheme.

Because of its subsampling operations in every subband, it is well known that the standard DWT[5] produces a shift dependent signal representation. In order to solve this problem, Rockinger presented a perfectly shift invariant wavelet fusion scheme called SIDWT[1,2,6]. However, this method is much computationally expensive due to its high redundancy ($2^m \times n : 1$ for m-D and n-level decomposition) of the representation.

In this paper, the decomposition of the image sequences was gotten by using the DT-CWT[8−10]. Not only can the DT-CWT attain approximate shift invariance and good directional selectivity, but also reduce much redundancy (its limited redundancy is $2^m : 1$ for m-D and any levels decomposition) compared with the SIDWT presented by Rockinger. Moreover, it can achieve the perfect reconstruction using short linear-phase filters. Figure 3 shows the decomposition and reconstruction (or its inverse transform) scheme of the DT-CWT for its one-dimension (1D). It can be easily extended to two-dimension (2D) by separably filtering along rows and columns.

A special fusion rule should be employed in object region to preserve the full information as much as possible in the target region. We assume that target detection gives M target maps: $T_{\rm IR} = \{t^1_{\rm IR}, t^2_{\rm IR}, \cdots, t^M_{\rm IR}\}$ in IR frame and N target region maps: $T_{\rm V} = \{t^1_{\rm V}, t^2_{\rm V}, \cdots, t^N_{\rm V}\}$ in the corresponding visible frame. The target maps ($T_{\rm J}$) in both source frames are analyzed jointly by $T_{\rm J} = T_{\rm IR} \cup T_{\rm V}$. Then the frame is segmented into three sets: single target region sets ($T_{\rm S}$), overlapping target region sets ($T_{\rm O}$), and background region set (B). Overlapped target regions can be denoted as $T_{\rm O} = T_{\rm IR} \cap T_{\rm V}$. Single target region sets are shown as $T_{\rm S} = T_{\rm J} \cap \bar{T}_{\rm O}$, where no targets overlap. Clearly, $T_{\rm J} = T_{\rm S} \cup T_{\rm O}$. Background regions are $B = \bar{T}_{\rm J}$.

In the single target regions, fusion rule (that is coefficients selecting method) can be written as

$$c_{\rm f}(x,y) = \begin{cases} c_{\rm ir}(x,y), & \text{if } (x,y) \in T_{\rm IR} \\ c_{\rm v}(x,y), & \text{if } (x,y) \in T_{\rm V} \end{cases}. \tag{1}$$

In a connected overlapped target region $t \in T_{\rm O}$, a similarity measure between two source images is defined as

$$M(t) = \frac{2 \cdot \sum\limits_{(x,y)\in t} I_{\rm ir}(x,y) \cdot I_{\rm v}(x,y)}{\sum\limits_{(x,y)\in t} [I_{\rm ir}(x,y)]^2 + \sum\limits_{(x,y)\in t} [I_{\rm v}(x,y)]^2}, \tag{2}$$

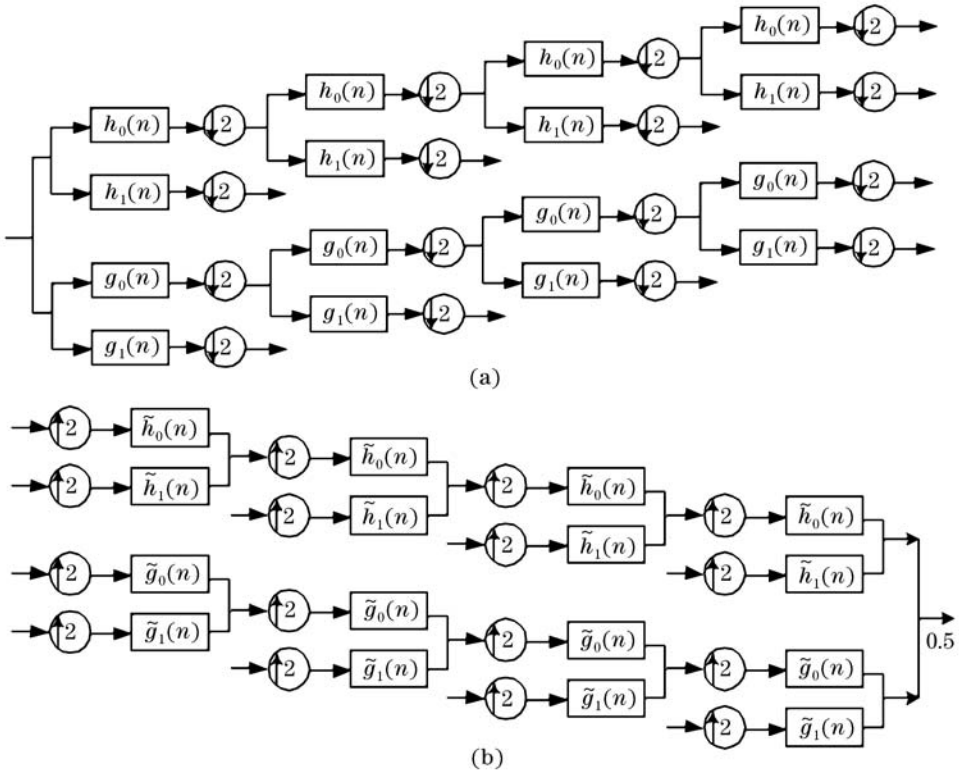

Fig. 3. DT-CWT and its inverse transform.

where $I_{\rm ir}$ and $I_{\rm v}$ denote the IR and visible frames, respectively. Then an energy index of the coefficients within the overlapped region is computed respectively in IR and visible frames,

$$S_i(t) = \sum_{(x,y)\in t} c_i(x,y)^2, \tag{3}$$

where $t \in T_{\rm O}$ and $i = {\rm ir}, {\rm v}$ means the IR and visible frames respectively. A threshold of similarity α is introduced where $\alpha \in [0,1]$ and normally $\alpha = 0.85$ (it is selected from many experimental results to get the best one) is appropriate.

In case $M(t) < \alpha$ the fusion rule (that is coefficients selecting method) in overlapping target region $t \in T_{\rm O}$ can be written as

$$c_{\rm f}(x,y) = \begin{cases} c_{\rm ir}(x,y), & \text{if } S_{\rm ir}(t) \geq S_{\rm v}(t) \\ c_{\rm v}(x,y), & \text{otherwise} \end{cases}. \tag{4}$$

In case $M(t) \geq \alpha$, a weight average method is adopted as

$$c_{\rm f}(x,y) = \begin{cases} w_{\max}(t)\cdot c_{\rm ir}(x,y) + w_{\min}(t)\cdot c_{\rm v}(x,y), \\ \qquad \text{if } S_{\rm ir}(t) \geq S_{\rm v}(t) \\ w_{\min}(t)\cdot c_{\rm ir}(x,y) + w_{\max}(t)\cdot c_{\rm v}(x,y), \\ \qquad \text{if } S_{\rm ir}(t) < S_{\rm v}(t) \end{cases}, \tag{5}$$

where the weights $w_{\min}(t)$ and $w_{\max}(t)$ can be obtained as

$$\begin{cases} w_{\min}(t) = \frac{1}{2}\left(1 - \frac{1-M(t)}{1-\alpha}\right) \\ w_{\max}(t) = 1 - w_{\min}(t) \end{cases}. \tag{6}$$

Finally, in the background regions, the simplest MS rule is adopted.

Here, the real world image sequences (dynamic images) were applied to experiment the fusion scheme proposed in this paper. The experimental steps are as follows: 1) Target detection: (a) segment the IR image with the fuzzy c-means clustering method; (b) a confidence measure for each candidate region is computed to get the target and the background regions. 2) Use the DT-CWT to transform the source image sequences in target and background regions. 3) Image fusion: (a) use the fusion rules proposed above to fuse the wavelet coefficient in the target region; (b) use the MS rule to fuse the wavelet coefficient in the background region. 4) Use the inverse dual-tree complex wavelet transform (DT-CWT) to get the fused image. 5) Use the model matching method to find the target regions in the subsequent frames and go to the step 2) until the last frame.

Figure 4 shows some results of the image sequences fusion as well as the spatially registered visible and IR (thermal 3—5 mm) source image sequences, which consist of 32 subsequent frames of each one. Figure 5 shows the corresponding inter-frame-difference (IFD) with images in Figs. 4(a) and (b). Directly observed, the results show that the proposed method of DT-CWT is superior to that of DWT when they all use the target detection scheme.

In order to quantify the temporal stability and consistency, we use the quantitative measure $I((S_{\rm v}, S_{\rm ir}), F)$, whose more details can be found in Ref. [1]. For every wavelet fusion method, we need to compute the average mutual information (AMI) instead of the 31 sets of the IFDs. The higher the AMI is, the better shift invariance, temporal stability and consistency the fusion method has. In the Table 1, DB4 means the Daubechies wavelets of

Table 1. AMI for the IFDs of Visual and IR Image Sequences

Fusion Method	DWT DB4	DWT BIOR4.4	DT-CWT Q-Shift9	DWF DB4	DWF BIOR4.4
Pixel-Based	1.6152	1.6208	2.0867	2.1435	2.1527
Feature-Based	1.7225	1.7263	2.2989	2.3147	2.3252

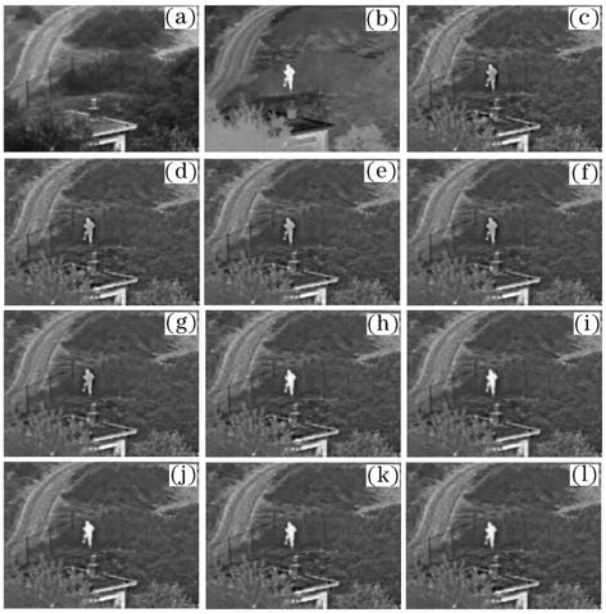

Fig. 4. (a) and (b) A pair of frames of visible and IR image sequences; (c) fused frame using the pixel-based scheme with DWT (DB4); (d) fused frame using the pixel-based scheme with DWT (BIOR4.4); (e) fused frame using the pixel-based scheme with DT-CWT (Q-shift9); (f) fused frame using the pixel-based scheme with DWF (DB4); (g) fused frame using the pixel-based scheme with DWF (BIOR4.4); (h)—(l) corresponding fused frames of (c)—(g) using the proposed feature-based dynamic fusion scheme.

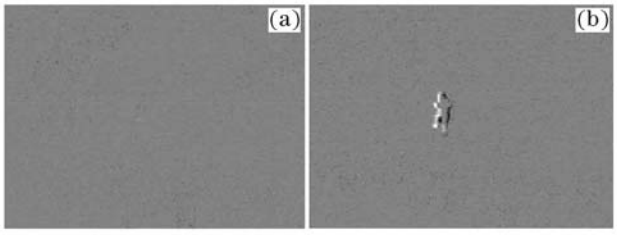

Fig. 5. (a) IFD between the current frame in Fig. 4(a) and its pre-frame in sequences; (b) IFD between the current frame in Fig. 4(b) and its pre-frame in sequences.

4 orders, and BIOR4.4 denotes the biorthogonal wavelets, where the two "4" are orders of synthesis and analysis filters respectively, and Q-shift9 was designed by Kingsbury[10]. "Pixel-based" and "Feature-based" mean that it is to use the pixel-based fusion scheme (fusion frames one by one) and the proposed feature-based dynamic fusion scheme. Obviously, the measured results in Table 1 are consistent with the conclusion of visual inspection. Moreover, the DT-CWT performs better than DWT, and it can be comparable with the DWF method in respect of the temporal stability and consistency with lower computational cost for that the DT-CWT has lower redundancy (2^m : 1 for DT-CWT, but $2^m \times n$: 1 for DWF).

A novel feature-based fusion of IR and visible image sequences using target detection is proposed in this paper. The results, which experimented by real IR and visible image sequences, show that the proposed feature-based fusion scheme of infrared and visible image sequences can achieve good shift invariance, temporal stability and consistency, lower computational cost as well as much target information, which can make the information of background clearer and help the monitor operators improve their spatial perception ability towards the target scene under complex background. However, when faced the situation of background change and many motional targets sheltered, the proposed scheme in this paper needs more research and improvement.

This work was jointly supported by the National Natural Science Foundation of China (No. 60375008), the 2010 Shanghai EXPO Special Project of National Key Technologies R&D Program (No. 2004BA908B07), and the Shanghai-NRC International Co-operate Project (No. 05SN07118). The original IR and visible images, which are available online at www.imagefusion.org, are kindly supplied by Alexander Toet of the TNO Human Factors Research Institute. C. Liu's e-mail address is liucongyi@sjtu.edu.cn.

References

1. O. Rockinger, IEEE Trans. Image Processing **3,** 288 (1997).
2. O. Rockinger and T. Fechner, Proc. SPIE **3374,** 378 (1998).
3. H. Wang, Z. Jing, and J. Li, Chin. Opt. Lett. **1,** 523 (2003).
4. Z. Li, Z. Jing, and S. Sun, Chin. Opt. Lett. **2,** 578 (2004).
5. S. G. Mallat, IEEE Trans. Pattern Analysis and Machine Intelligence **11,** 674 (1989).
6. O. Rockinger, in *Proceedings of 16th Leeds Annual Statistical Research Workshop* 149 (1996).
7. A. Yilmaz, K. Shafique, and M. Shah, Image and Vision Computing **21,** 623 (2003).
8. N. G. Kingsbury, in *Proceedings of 8th IEEE DSP Workshop* 86 (1998).
9. I. W. Selesnick, R. G. Baraniuk, and N. G. Kingsbury, IEEE Signal Processing Magazine **22,** 123 (2005).
10. N. G. Kingsbury, in *Proceedings of IEEE Int. Conf. Image Processing* **2,** 375 (2000).