This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2711422, IEEE Transactions on Multimedia

An Adaptive Fusion Algorithm for Visible and Infrared Videos Based on Entropy and the Cumulative Distribution of Gray Levels [1]

Hai-Miao Hu[1, 2, *], Jiawei Wu[1, *], Bo Li[1, 2, #], Qiang Guo[1], Jin Zheng[1, 2]
1 Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, Beijing 100191, China
2 State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China
* The authors contributed equally to this work.
# The corresponding author: boli@buaa.edu.cn

Abstract: Visible videos captured under different weather conditions may exhibit different characteristics, and thermal infrared videos are easily affected by ambient temperature variations; this sensitivity to environmental conditions makes the fusion of visible and thermal infrared videos a challenge. This paper proposes an adaptive fusion algorithm for visible and infrared videos, and uses cumulative distribution of gray levels and the entropy to adaptively retain infrared-hot targets and visible textures. The original visible and infrared frames are decomposed into two layers, namely, the base layer and the detail layer. The guided filter is employed to decompose frames due to its high efficiency. Two weight maps, one for the infrared base layer and one for the visible base layer, are adaptively generated based on the cumulative distribution of gray levels and the entropy, respectively. The visible base layer and the infrared base layer are fused based on their weight maps. The final fusion result is obtained by combining the fused base layer with the visible detail layer. Experimental results demonstrate that the proposed algorithm can achieve better fusion results compared with state-of-the-art methods.

Index Terms— visible video, infrared video, adaptive video fusion, cumulative distribution function, entropy

1. Introduction

Both thermal infrared videos and visible videos are widely used in many practical applications, especially for video surveillance. Visible videos often have rich textures; whereas the targets may not be clear because of camouflage, including hidden in the dark, forest, or smoke, etc. In these conditions, the targets cannot be seen clearly. Infrared videos can see the hidden targets; whereas infrared videos usually lack of textures. Since they have different complementary advantages, visible and thermal infrared videos can be fused to improve the overall perceptual quality [1,2]. However, visible videos captured under different weather conditions (e.g., on a foggy day or during the night) may exhibit different characteristics, and thermal infrared videos are easily affected by ambient temperature variations; this sensitivity to environmental conditions makes the fusion of visible and thermal infrared videos a challenge.

A large number of fusion approaches for visible and infrared videos have been proposed [3-15]. They can be divided into two categories, namely, video-based fusion approaches [3,4] and frame-based fusion approaches [5-15].

Video-based approaches use three-dimensional (3D) non-separate multi-scale transform (MST) tools. Zhang et al use the 3D surfacelet transform (3D-ST) [3] and the 3D uniform discrete curvelet transform (3D-UDCT) [4] to fuse videos. In such 3D algorithms, video signals are treated as existing in a special set of three dimensions (i.e., two spatial dimensions and one temporal dimensions), and the signals are fused simultaneously. However, these 3D transforms have a high computational complexity and require the complete information of the entire video, which is only feasible for off-line applications.

In frame-based approaches, two videos are fused frame by frame, which can be done in either the spatial domain or the transform domain; these methods include Karhunen-Loève transform fusion [5], Laplacian pyramid fusion [6], wavelet fusion [7], curvelet fusion [8], contourlet fusion [9,10], generalized random walk fusion [11], and Markov random field fusion [12]. Wang et al [13] use a genetic algorithm to perform adaptive fusion, but a genetic algorithm requires multiple iterations and thus is inefficient. Li et al[14] consider the time complexity and propose a method of image fusion using a guided filter. However, this algorithm ignores the differences between visible and infrared frames and generates saliency maps of the visible and infrared images using the same algorithm, which can easily result in information loss. Ma et al [15] consider the differences between visible and infrared frames and propose a fusion algorithm based on the minimization of the total variation (TV). However, this algorithm tends to retain the thermal radiation information, which will affect the global performance of the fusion result.

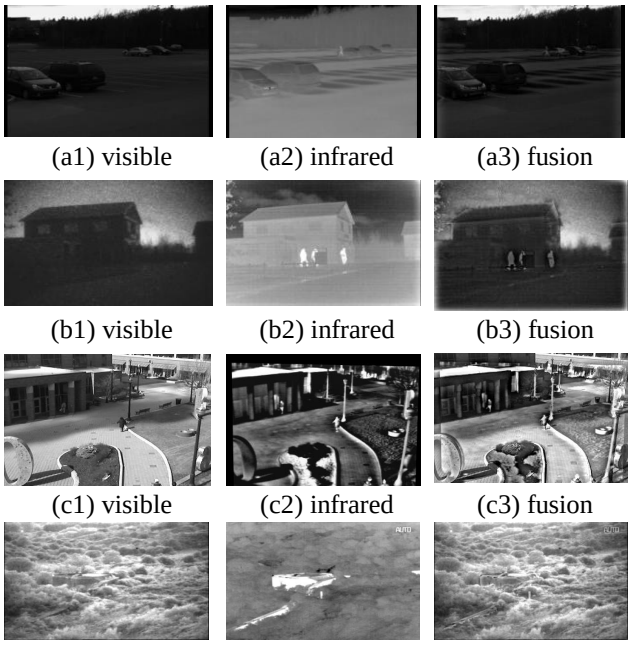
Moreover, the fusion algorithms discussed above focus on preserving the details of the images as well as maximizing the textures of the sources. However, they fail to account for the different characteristics of different kinds of videos acquired under different environmental conditions, which will result in the following two problems.

Firstly, visible videos captured under different weather conditions, such as on a foggy day or during the night, may exhibit different characteristics due to environmental interference. For example, the high-frequency component of a visible video captured during the daytime contains texture information, whereas for a visible video captured at night, the high-frequency component usually contains noise because of the lack of light. Fig. 1 shows a comparison of visible videos captured during the daytime and at night. The corresponding Laplacian pyramid fusion results are also shown in Fig. 1. As shown in this figure,

[1] This work was partially supported by the National Key Research and Development Plan (Grant No. 2016YFC0801003), the National Natural Science Foundation of China (No.61370121, 61421003).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2711422, IEEE Transactions on Multimedia

the fusion algorithm is easily influenced by the quality of the visible frames. When a visible frame shows a poor image quality and is globally dark, the fusion result will also be of poor quality. Therefore, the quality of the visible videos should be considered during the selection of fusion strategies, and an adaptive fusion algorithm should be implemented that considers the unique characteristics of visible videos.



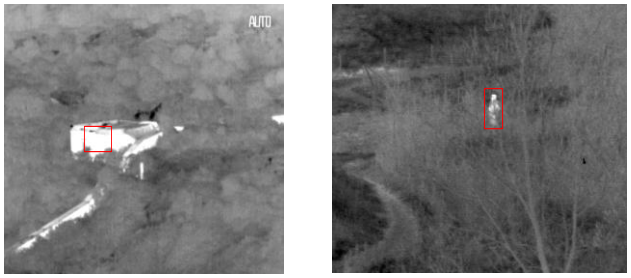
(a1) visible (a2) infrared (a3) fusion

(b1) visible (b2) infrared (b3) fusion

(c1) visible (c2) infrared (c3) fusion

(d1) visible (d2) infrared (d3) fusion

Fig. 1 Examples of visible videos captured under different conditions. ((a) and (b) were captured at night, whereas (c) and (d) were captured during the daytime.)



(a) (b)

Fig. 2 Gray-level comparison of hot targets captured in different environments. (a) shows an infrared video captured in a hot environment, whereas (b) shows an infrared video captured in a cold environment.

Secondly, in thermal infrared videos, hot objects are represented in the low-frequency component, and the texture information contained in such a video is useless. However, the low-frequency component of thermal infrared videos is easily affected by ambient temperature variations, which can affect how well hot objects are preserved. Fig. 2 shows a comparison between images of hot objects in a cold environment and hot objects in a hot environment. As shown in Fig. 2, a gray level of 110 can be considered to represent a hot target in Fig. 2(b), whereas it can only be treated as background in Fig. 2(a). The gray levels of the hot regions in Fig. 2(a) are as high as 200. For the effective extraction of hot objects, the ambient temperature should be considered in the fusion algorithm, and an adaptive adjustment that considers this characteristic of infrared videos is important for ensuring the quality of the final fusion results.

Therefore, this paper proposes an adaptive fusion algorithm for visible and infrared videos that maintains both infrared-hot objects and visible textures. First, the original visible and infrared frames are decomposed into two layers, namely, the base layer and the detail layer, using a guided filter. The base layer mainly contains overall variations in intensity, whereas the detail layer contains details. Second, separate weight maps are adaptively generated for the visible base layer and the infrared base layer. For the visible video, the weight map is adaptively generated based on the entropy information of different frames. For the infrared video, the weight map is adjusted based on the cumulative distribution function (CDF) of the gray levels to globally suppress the interference from the environmental temperature while emphasizing hot objects. Finally, the visible base layer and the infrared base layer are fused based on their weight maps, and the final fusion result is obtained by combining the fused base layer with the visible detail layer. Experimental results demonstrate that the proposed algorithm not only can achieve adaptive fusion without the loss of either infrared-hot objects or visible textures but also can achieve better fusion results compared with state-of-the-art fusion methods.

Note that the detail layer in the infrared video is not used in the proposed algorithm because the majority of the useful information in an infrared video (e.g., hot objects) is contained in the base layer, whereas the detail layer of an infrared video usually contains noise. The remainder of this paper is organized as follows. The guided filter, the cumulative distribution function and the entropy used in the proposed algorithm are described in Section 2. Section 3 discusses the proposed fusion algorithm. Section 4 presents the experimental results. Some discussion and an extension of the proposed algorithm are presented in Section 5. Finally, the paper is concluded in Section 6.

2. Preliminaries

This section describes the concepts on which the proposed fusion algorithm is based. These concepts, including the guided filter, the cumulative distribution function and the entropy, are described as follows.

A. Guided filter

A guided filter [16] is a kind of filter that is widely applied in computer vision and graphics[17,18]. Compared with a bilateral filter, it exhibits a beneficial property of edge-preserving smoothing while being efficient to compute (specifically, the calculation complexity is O(N), regardless of the kernel size and the intensity range); therefore, it is a suitable choice for video fusion.

The theory underlying the use of the guided filter is introduced as follows. Suppose that the filter output q is a linear transform of the guidance image I in a square window ω_k of size $(2r+1)\times(2r+1)$ centered on pixel k:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2711422, IEEE Transactions on Multimedia

$$q_i = a_k I_i + b_k, \forall i \in \omega_k \quad (1)$$

where (a_k, b_k) are linear coefficients that are assumed to be constant in ω_k. To minimize the difference between *q* and the input image *p*, a cost function in the window ω_k is built as follows:

$$E(a_k, b_k) = \sum_{i \in \omega_k} ((a_k I_i + b_k - p_i)^2 + \varepsilon a_k^2) \quad (2)$$

Equation (2) is the linear ridge regression model, and its solution is given by

$$a_k = \frac{\frac{1}{|\omega|}\sum_{i \in \omega_k} I_i p_i - \mu_k \bar{p}_k}{\sigma_k^2 + \varepsilon} \quad (3)$$

$$b_k = \bar{p}_k - a_k \mu_k \quad (4)$$

where μ_k and σ_k^2 are the mean and variance of I in ω_k, respectively; $|\omega|$ is the number of pixels in ω_k; and $\bar{p}_k = \frac{1}{|\omega|}\sum_{i \in \omega_k} p_i$ is the mean of *p* in ω_k. Finally, the filter output is computed as follows:

$$q_i = \bar{a}_i I_i + \bar{b}_i \quad (5)$$

Here, $\bar{a}_i = \frac{1}{|\omega|}\sum_{k \in \omega_i} a_k$ and $\bar{b}_i = \frac{1}{|\omega|}\sum_{k \in \omega_i} b_k$ are the average coefficients of all windows overlapping *i*.

Guided filters can be applied in many situations, such as for detail enhancement, denoising, or guided feathering. In this paper, we treat the filter only as an edge-preserving smoothing operator, which is based on a special case in which the guidance image *I* is identical to the filter input *p*. If $I \equiv p$, we can conclude from (3) and (4) that $a_k = \sigma_k^2 / (\sigma_k^2 + \varepsilon)$ and $b_k = (1 - a_k)\mu_k$. In the high-variance regime, we have $\sigma_k^2 >> \varepsilon$ and, consequently, $a_k \approx 1$ and $b_k \approx 0$, whereas in the flat regime, we have $\sigma_k^2 << \varepsilon$ and, consequently, $a_k \approx 0$ and $b_k \approx \mu_k$. Therefore, given an input image *p*, its edge-preserving smoothed output is treated as the base layer, whereas the difference between the input and the base layer is the detail layer.

B. Cumulative distribution function

In probability theory and statistics, the cumulative distribution function of a real-valued random variable *X* evaluated at *x* is the probability that *X* will take a value less than or equal to *x*:

$$CDF_X(x) = P(X \le x) \quad (6)$$

where $P(X \le x)$ is the probability that the random variable *X* will take a value less than or equal to *x*.

In this paper, the CDF is used to describe the ambient temperature in an infrared video. More specifically, considering that a hot target in a cold environment may not correspond to a high absolute gray level, the CDF is applied to enable the identification of hot targets of this kind; the use of the CDF ensures that a hot target will have a high weight regardless of the ambient temperature.

C. Entropy

In this paper, we need a metric to describe the quality of a visible or infrared video. Many metrics have been proposed to addressed this problem, such as those introduced in [19] and [20]. In this paper, we use the Shannon entropy to describe the image quality. The Shannon entropy (EN) is defined as follows:

$$EN = -\sum_{i=0}^{L-1} p_F(l) \log_2 p_F(l) \quad (7)$$

where L is the number of gray levels and $p_F(l)$ is the normalized histogram of the image.

Entropy can be described as a measure of the amount of disorder in a system. In the case of an image, a low-entropy image exhibits a small number of gray levels, whereas a high-entropy image exhibits a large number of gray levels. In a structured image, centralized gray levels correspond to dark or overexposed illumination, which implies poor image quality, whereas balanced gray levels correspond to moderate illumination, which implies good image quality.

3. Summary of the proposed algorithm

The proposed algorithm consists of three main steps, namely, frame decomposition using the guided filter, adaptive base layer fusion, and result generation. A block diagram of the proposed algorithm is shown in Fig. 3. The three steps are further elaborated in the following sections.

A. Frame decomposition using the guided filter

First, the visible and infrared frames are both decomposed into a base layer and a detail layer using the guided filter. The base layer is calculated as follows:

$$B^{vis} = \mathrm{GF}_{r,\varepsilon}(I^{vis}, I^{vis}) \quad (8)$$

$$B^{inf} = \mathrm{GF}_{r,\varepsilon}(I^{inf}, I^{inf}) \quad (9)$$

where I^{vis} or I^{inf} represents the original visible or infrared frame, respectively; *r* and *ε* are the parameters of the guided filter; GF is the guided filter function; and B^{vis} or B^{inf} represents the base layer of the visible or infrared frame, respectively.

Once the base layer has been obtained, the detail layer *D* can be generated:

$$D^{vis} = I^{vis} - B^{vis} \quad (10)$$

$$D^{inf} = I^{inf} - B^{inf} \quad (11)$$

This two-layer decomposition is applied to separate the base layer from the detail layer such that when the base

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2711422, IEEE Transactions on Multimedia

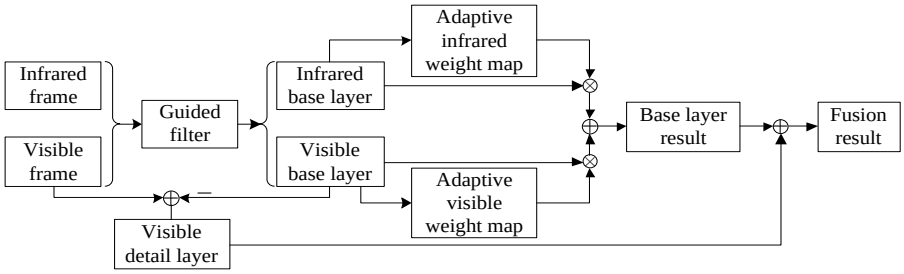


Fig. 3 Block diagram of the proposed algorithm

layer is fused, the infrared textures will not affect the real textures, namely, the visible textures. Moreover, infrared-hot objects will always appear in the base layer.

B. Base layer fusion based on weight maps

Adaptive generation of the infrared weight map based on the cumulative distribution function of the gray levels

Hot objects, which are the most important information contained in infrared frames, need to be emphasized in the fusion result. For most infrared sensors, a higher pixel value indicates a hotter region, whereas a lower pixel value indicates a colder region. In some papers, such as [21], the absolute pixel value is used to evaluate the weight of each pixel; thus, the evaluation is not adaptive in most cases. On a cold winter night, the pixel values of all objects will be reduced because of the low temperature. Hot objects will still have relatively high pixel values compared with other objects, but their absolute values will not be high in general.

Considering this phenomenon, an adaptive infrared weight map algorithm is proposed in this paper based on the cumulative distribution function of the gray levels. For a pixel value *k*, the CDF of *k* in frame *f* can be calculated as follows:

$$CDF_f(k) = P(x \le k) \tag{12}$$

where $P(x \le k)$ represents the probability that a pixel *x* in *f* will have a value of less than or equal to *k*.

From this equation, it is clear that the CDF of the gray levels is adaptive with respect to the environmental temperature. If $CDF_f(k) \approx 1$, then *k* must be a relatively high pixel value in this particular frame.

Using this definition of the CDF of the gray levels, an adaptive infrared map can be defined as follows:

$$inf\_w_{i,j} = \begin{cases} CDF_{B^{inf}}(B_{i,j}^{inf}) & CDF_{B^{inf}}(B_{i,j}^{inf}) > 0.9 \\ CDF_{B^{inf}}(B_{i,j}^{inf}) / s & CDF_{B^{inf}}(B_{i,j}^{inf}) <= 0.9 \end{cases} \tag{13}$$

where *i* and *j* are the pixel indices. If $CDF_{B^{inf}}(B_{i,j}^{inf}) \le 0.9$, then a punishment parameter *s* is applied. If *s* is too large, then hole artifacts will appear in the result, as shown in section 5.A. If *s* is too small, then regions that are not sufficiently hot will affect the fusion result. In this paper, *s* is set to 2 by default.

Since hot objects occupy only a small portion of the total frame in most cases, especially in video surveillance applications, a rather high threshold can be expected to be appropriate, and experimental results indicate that 0.9 is a suitable value. Therefore, the threshold is set to 0.9 in this paper.

Adaptive generation of the visible weight map based on the entropy

Visible textures, which are the most valuable information in a visible frame, should be preserved in the fusion result. However, poor infrared textures can affect these valuable visible textures. Therefore, a higher weight should be assigned to the textures in the visible frame.

To ensure that the visible textures are retained, the weight map vis_w is calculated as follows:

$$H = I^{vis} * L \tag{14}$$

$$vis\_w_{i,j} = abs(H_{i,j}) / \max(H) \tag{15}$$

where *L* is a 3×3 Laplacian operator. The divisor max(*H*) normalizes the visible weight map to the interval of [0,1].

Since both the infrared weight map and the visible weight map are normalized to [0,1], they are considered to be equally important. However, in most cases, they are not of strictly equal importance. In challenging environments, such as foggy and dark environments, visible frames are unlikely to be of good image quality and may even be noisy, whereas under good conditions, the visible textures will be rich and contain considerable information.

Therefore, to endow the visible weight map with adaptive properties, the image quality should be used to modify the range of the visible weights. In this paper, the entropy is used to represent the image quality, and the visible weight range is modified as follows:

$$vis\_w'_{i,j} = \begin{cases} vis\_w_{i,j} \times (1 + e_v - e_i) & e_v > e_i \\ vis\_w_{i,j} / (1 + e_i - e_v) & e_i < e_v \\ vis\_w_{i,j} & e_v = e_i \end{cases} \tag{16}$$

where e_v and e_i represent the entropy values of the visible and infrared frames.

When the visible frame is of poor image quality, the range of the visible weights will be shrunk to reduce the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2711422, IEEE Transactions on Multimedia

influence of noise. By contrast, when the image quality of the visible frame is good, the range of the visible weights will be expanded to emphasize the visible textures.

Base layer fusion

Finally, the base layer of the fusion result is generated. It should be noted that the base layer of the fusion result will be primarily based on the visible frame, not the infrared frame. Therefore, the base layer is calculated as follows:

$$B_{i,j}^{fuse} = \frac{B_{i,j}^{vis} + B_{i,j}^{vis} \times vis\_w_{i,j}^{'} + B_{i,j}^{inf} \times inf\_w_{i,j}}{1 + vis\_w_{i,j} + inf\_w_{i,j}} \quad (17)$$

C. Fusion result generation

Once the fused base layer has been generated, the final result can be obtained by combining the fused base layer with the visible detail layer:

$$I^{fuse} = B^{fuse} + d \times D^{vis} \quad (18)$$

where d is a parameter that controls the level of detail added to the final result; in this paper, d is set to 1 by default.

Notably, the infrared detail layer is simply discarded without being used because in the infrared frame, the only information of interest is the hot objects, which appear only in the infrared base layer.

In fact, if one is not concerned about introducing artifacts, then the infrared frame can be converted into a binary frame in which all details are discarded. However, this paper uses a base layer instead of a binary frame to avoid the generation of artifacts.

4. Results and discussion

This section describes five experiments conducted based on four datasets and presents an objective quality evaluation comparing the proposed algorithm with state-of-the-art fusion methods.

To verify the advantages of our fusion algorithm, experiments were performed on four datasets. The first dataset is OTCBVS [22]. There are 12 sub-datasets in OTCBVS, of which dataset 03, the OSU Color-Thermal Database, was used in this study. This dataset was obtained by simultaneously recording visible and infrared videos at busy pathway intersections on the Ohio State University campus. The infrared sensor used was a Raytheon PalmIR 250D with a 25 mm lens, and the visible sensor was a Sony TRV87 Handycam. The image size is 320x240 pixels. The fusion result for a frame from OTCBVS is presented in Fig. 4.

The second dataset is the TNO Image Fusion Dataset [23]. The TNO Image Fusion Dataset contains multispectral (intensified visual, near-infrared, and long-wave infrared or thermal) nighttime imagery of various military-relevant scenarios, captured using various multiband camera systems. Fig. 5 and Fig. 6 present fusion results from the TNO dataset. All images have dimensions of 768x576 pixels.

The third dataset is the INO image fusion dataset [24]. This dataset is provided by the National Optics Institute of Canada and contains several pairs of visible/infrared videos representing different scenarios captured under different weather conditions. In this paper, Fig. 7 shows an image from this dataset that was shot during the evening at a park. The frame size is 328x254.

The fourth dataset is the Eden Project Multi-Sensor Dataset [25]. This dataset contains several pairs of videos, most of which correspond to camouflage scenarios. Fig. 8 shows fusion results for the Eden dataset, where the frame size is 560x468.

A. Experiments

To evaluate the performance of the proposed fusion algorithm, five experiments were performed on the four databases described above. The proposed fusion algorithm was compared with five other algorithms, namely, Laplacian pyramid fusion (LAP) [26], discrete wavelet transform fusion (DWT) [27], principle component analysis fusion (PCA), generalized random walk fusion (GRW) [10] , guided filter fusion (GFF) [13] and the proposed algorithm[2].

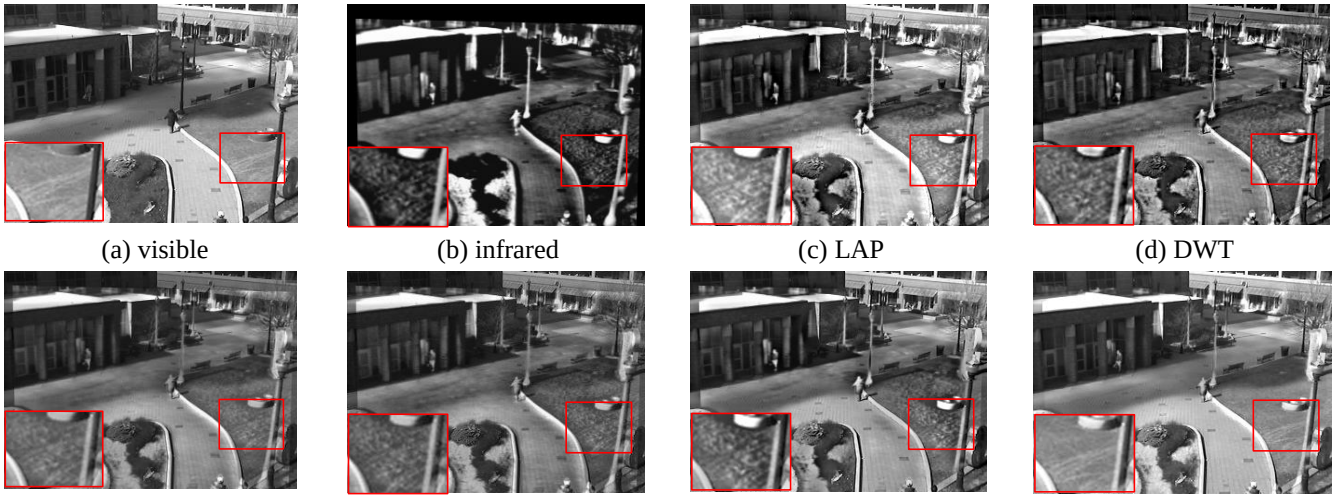
Fig. 4 shows the fusion results for a frame from the OTCBVS database. In addition, a close-up view of a smaller region is presented in the bottom left corner of each image. In the visible frame, the person hidden in the shadow is nearly invisible, whereas it can be clearly seen in the infrared frame. In addition, in the original infrared frame, some spots can be seen on the grass. These spots are temperature textures that exist only in the infrared frame and should not be retained in the fusion results. The various state-of-the-art algorithms treat these spots as important information and retain them in the fusion results. Because of these infrared spots, the real visible textures of the grass are obscured (as shown in Fig. 4(c)-(h)). In the result obtained using the proposed algorithm, the visible textures of the grass are successfully retained. Moreover, the original visible frame is of good quality, and only one hot object is not clear. Therefore, the fusion result should not include too much information from the infrared regions, and only the hot objects should be added to the visible frame. From the results, it can be seen that all fused images, except ours, contain excess information from the infrared regions.

Fig. 5 shows the fusion results for an image of a bunker. The original visible frame is full of visible textures, whereas the infrared frame contains invisible hot objects. The fusion results of the LAP, PCA, GRW and GFF algorithms fail to preserve both the infrared-hot objects and the visible textures. The results of DWT and our proposed algorithm successfully retain both types of information. However, the DWT result is affected by information from the infrared frame and has poor illumination properties. Only the result of our proposed algorithm shows the proper illumination and contains all of the desired information.

Fig. 6 shows the fusion results for an image of a shooter behind smoke. The visible frame contains a lot of smoke and the target is hard to identify, whereas the target can be seen clearly in the infrared frame. The LAP and PCA algorithms ignore the information in the visible frame, and the results are essentially identical to the infrared frame. The results of the GRW and GFF algorithms do consider the visible frame. However, they contain

[2] https://github.com/TiddyWu/An-Adaptive-Fusion-Algorithm-for-Visible-and-Infrared-Videos-Based-on-Entropy-and-the-Cumulative-Dis.

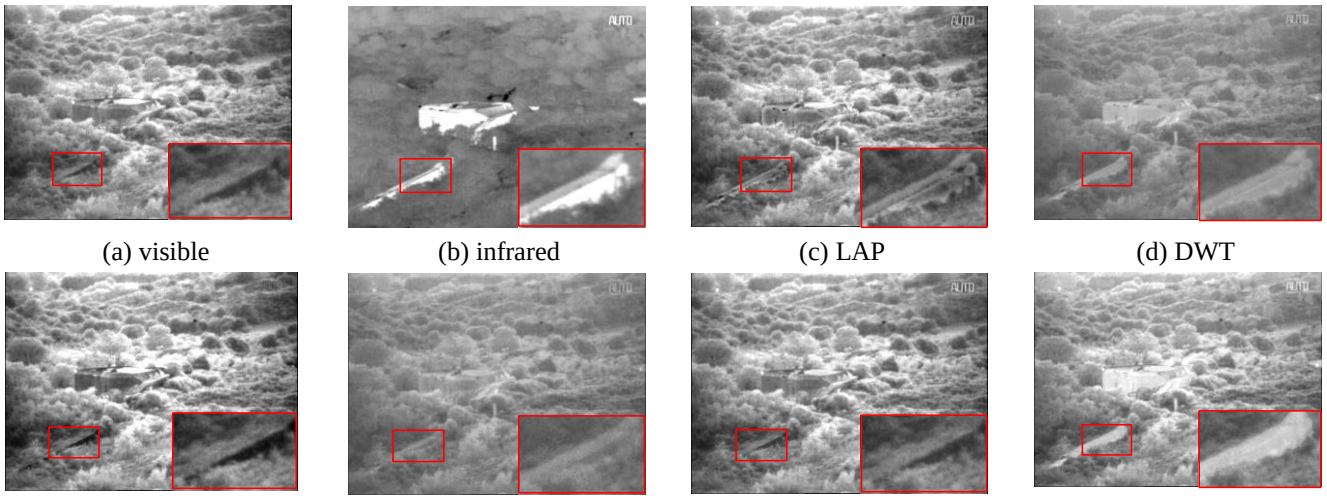
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2711422, IEEE Transactions on Multimedia



(a) visible (b) infrared (c) LAP (d) DWT

(e) PCA (f) GRW (g) GFF (h) Our algorithm

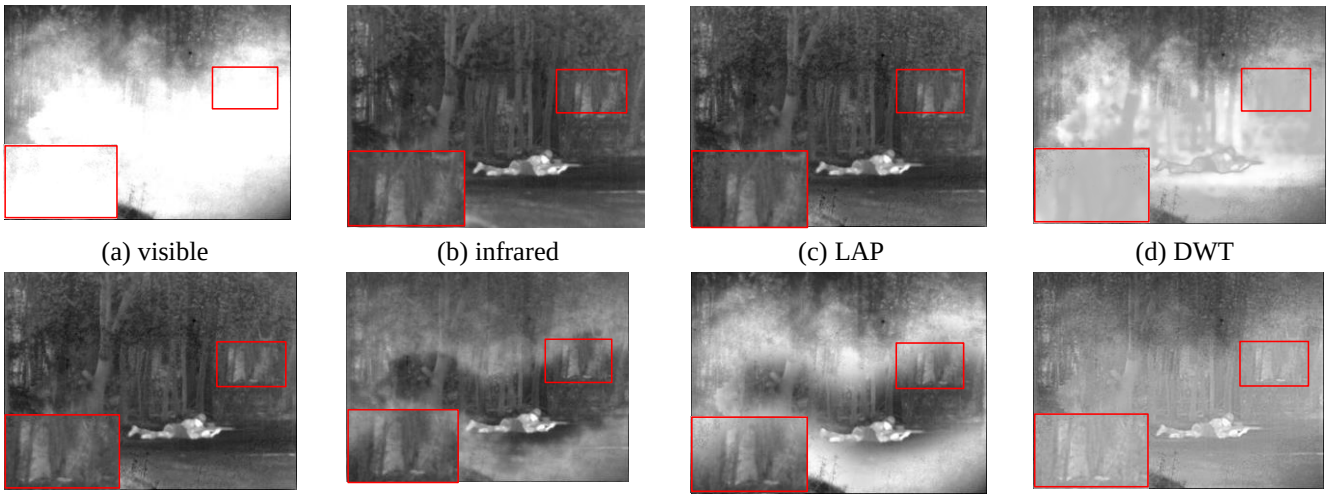
Fig. 4 Performance comparison of various fusion algorithms on the OTCBVS database.



(a) visible (b) infrared (c) LAP (d) DWT

(e) PCA (f) GRW (g) GFF (h) Our algorithm

Fig. 5 Performance comparison of various fusion algorithms on the TNO dataset (example 1).

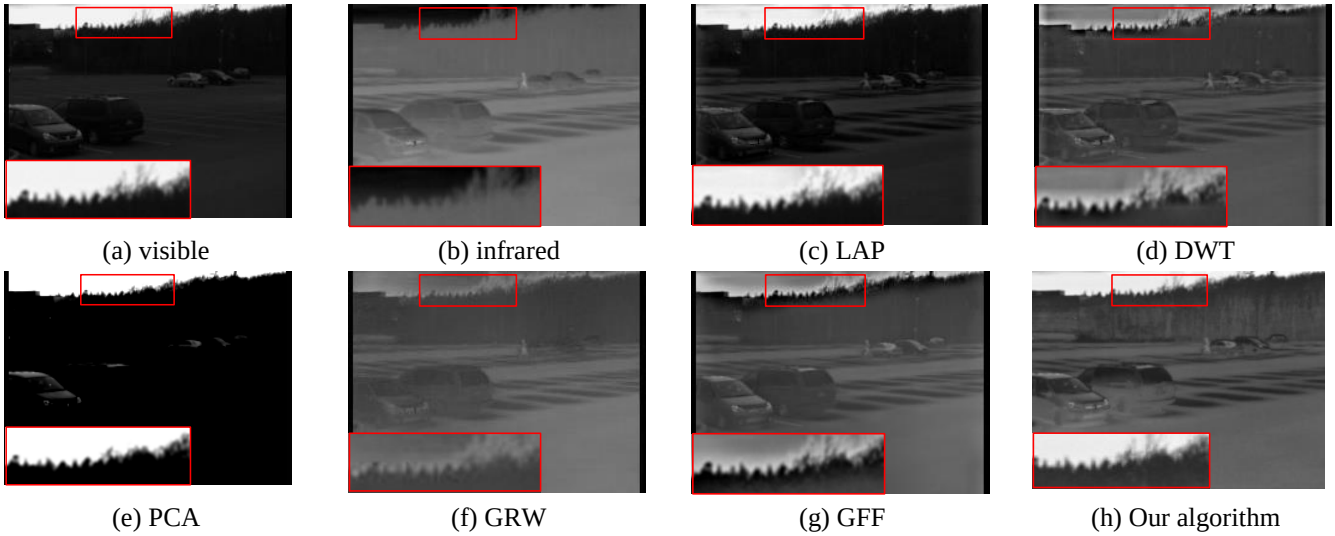


(a) visible (b) infrared (c) LAP (d) DWT

(e) PCA (f) GRW (g) GFF (h) Our algorithm

Fig. 6 Performance comparison of various fusion algorithms on the TNO dataset (example 2).

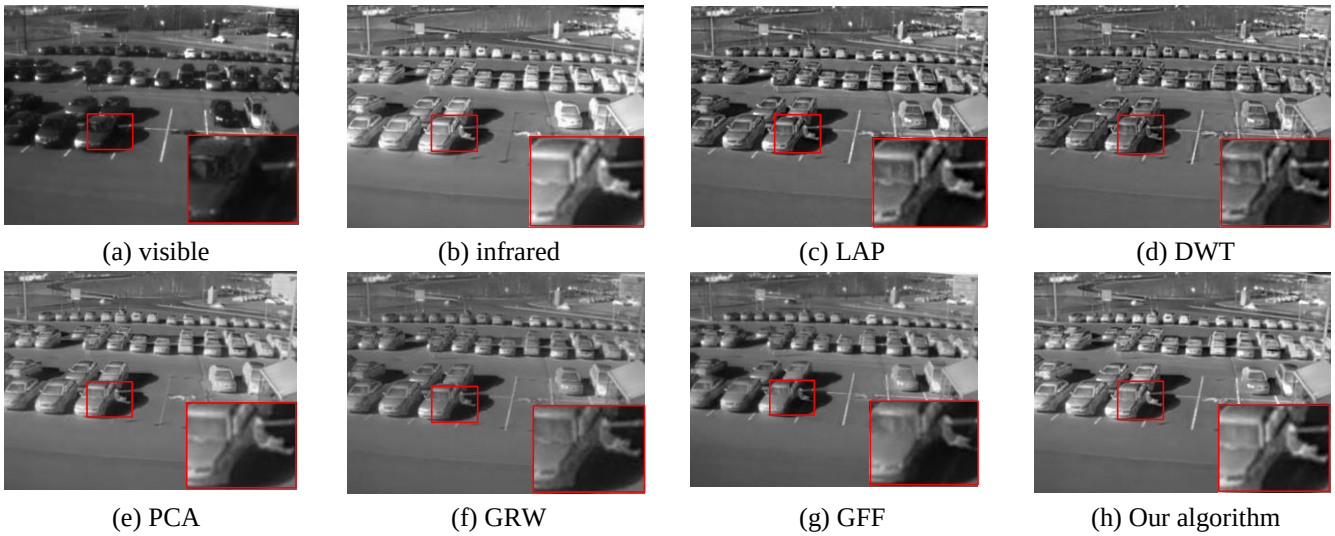
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2711422, IEEE Transactions on Multimedia



(a) visible (b) infrared (c) LAP (d) DWT

(e) PCA (f) GRW (g) GFF (h) Our algorithm

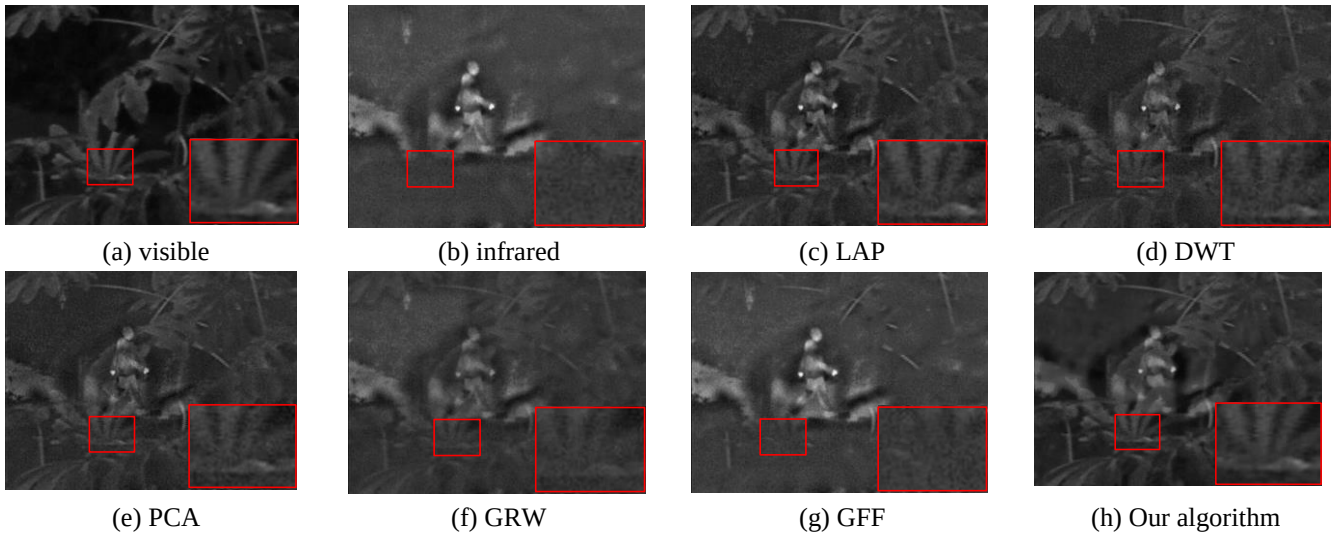
Fig. 7 Performance comparison of various fusion algorithms on the INO dataset(example 1).



(a) visible (b) infrared (c) LAP (d) DWT

(e) PCA (f) GRW (g) GFF (h) Our algorithm

Fig. 8 Performance comparison of various fusion algorithms on the INO dataset (example 2).

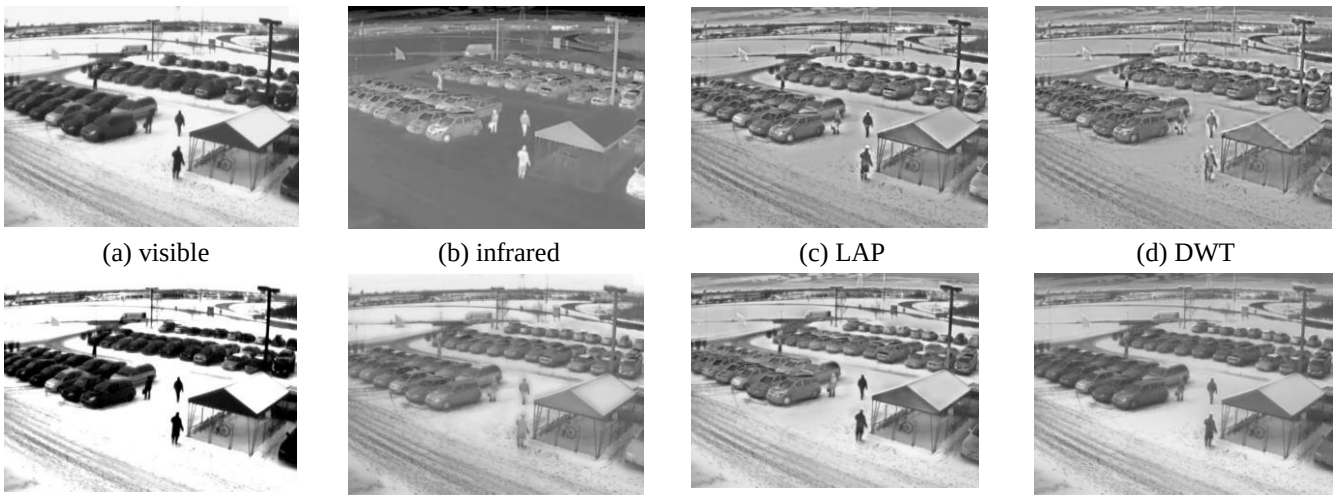


(a) visible (b) infrared (c) LAP (d) DWT

(e) PCA (f) GRW (g) GFF (h) Our algorithm

Fig. 9 Performance comparison of various fusion algorithms on the Eden dataset.

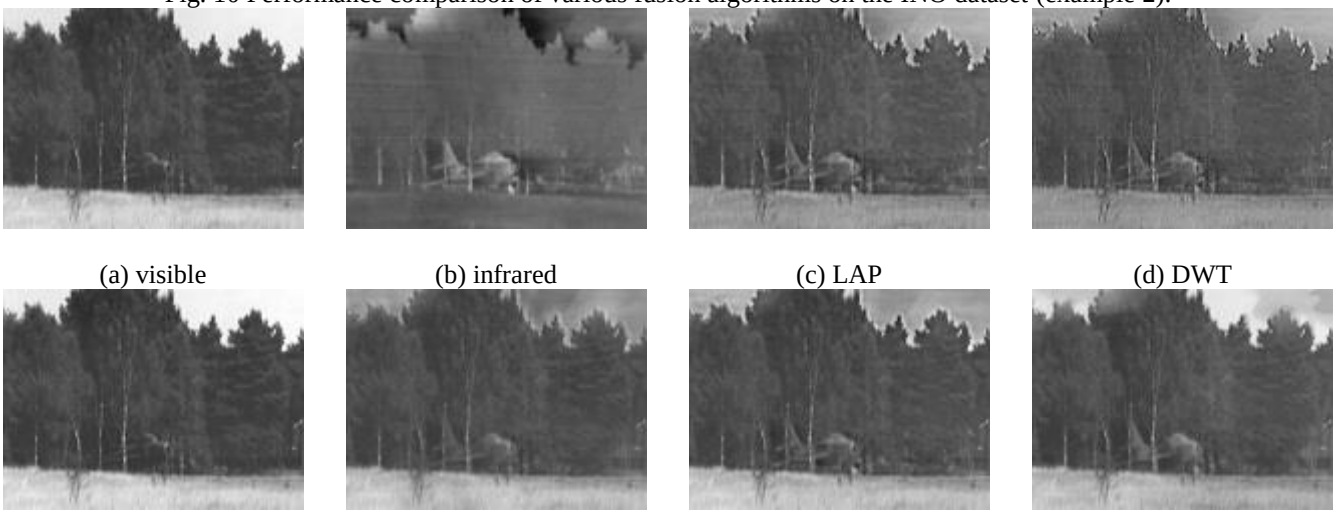
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2711422, IEEE Transactions on Multimedia



(a) visible (b) infrared (c) LAP (d) DWT

(e) PCA (f) GRW (g) GFF (h) Our algorithm

Fig. 10 Performance comparison of various fusion algorithms on the INO dataset (example 2).



(a) visible (b) infrared (c) LAP (d) DWT

(e) PCA (f) GRW (g) GFF (h) Our algorithm

Fig. 11 Performance comparison of various fusion algorithms on the TNO dataset (example 3).

severe hole artifacts, and the smoke around the shooter is missing. DWT and the proposed algorithm avoid the generation of artifacts, and both the smoke and the shooter are clear in the results. However, compared with the DWT result, the result obtained using our proposed algorithm has a better resolution.

Fig. 7 shows the fusion results for an image of a parking lot during the evening. Because of the poor illumination conditions, the image quality of the visible frame is poor, whereas that of the infrared frame is better. The LAP and PCA results look similar to the visible frame, with poor illumination. DWT, GRW, GFF and the proposed algorithm yield similar results. However, it is clear that the results of the DWT, GRW and GFF algorithms exhibit some "halo artifacts" at the top of the forest region. The LAP and PCA algorithms result in poor illumination, whereas our proposed algorithm performs the best.

Fig. 8 shows the fusion results of a parking lot. The visible frame contains rich textures, whereas the infrared frame contains poor textures, such as white lines on the ground. However, the visible frame cannot show the man sitting in the shadow, whereas the infrared frame can show it clearly. In all of the fusion results, the man sitting in the shadow can be retained, and the proposed algorithm can give the best result, with both rich textures and the hidden target.

Fig. 9 shows the fusion results for an image of a person walking behind some plants at night. The visible frame contains textures of the trees, whereas the infrared frame contains poor textures. However, the infrared frame includes invisible hot objects. The hot objects are properly emphasized in all results. However, from the close-up view of the infrared frame, it is clear that the infrared frame contains obvious noise. In nearly all fusion results except ours, the visible textures of the plants are affected by the infrared noise. Moreover, most of the visible textures are lost in the LAP, GRW and GFF results, while the result of the proposed algorithm contains nearly the same

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2711422, IEEE Transactions on Multimedia

textures as are present in the visible frame. Fig. 10 shows another pair of experiment from INO dataset, and the proposed algorithm achieves appropriate result.

Since some thresholds of the proposed algorithm are pre-defined according to the experiments and observations, the proposed algorithm may not outperform other algorithms in some cases. It can be observed from Fig.11 that the proposed algorithm can achieve comparative, or even better results when compared with other algorithms, but it cannot achieve a better result when compared with GFF. According to our result shown in Fig.11(h), part of the airplane lost texture in infrared frame. Note that the threshold can be optimally determined (e.g., through machine learning method), which can further improve the performance of the proposed algorithm and will be one of our ongoing works.

B. Objective Quality Evaluation

To objectively evaluate the performance of the proposed fusion algorithm, five objective fusion quality measures were adopted, namely, the global standard deviation, the information entropy, the gradient-based index $Q^{AB/F}$ [28], the mutual information (MI) [29], and the structural-similarity-based metric (SSIM) [30]. The default parameters were used for all objective measures.

(1) The global standard deviation (SD) is defined as follows:

$$SD = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} I_{i,j} - \bar{I}} \qquad (19)$$

where I is an image with dimensions of $M \times N$ and $\bar{I}$ is the average gray value of image I. The standard deviation reflects the extent to which the values of individual pixels in the image differ from the average value, which is also generally considered to correspond to the high-frequency information in an image. A higher standard deviation often indicates a better image quality.

(2) The gradient-based index $Q^{AB/F}$ is defined as follows:

$$Q^{AB/F} = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M} Q^{AF}(n,m)\omega^{A}(n,m) + Q^{BF}(n,m)\omega^{B}(n,m)}{\sum_{i=1}^{N}\sum_{j=1}^{M}(\omega^{A}(i,j)+\omega^{B}(i,j))} \qquad (20)$$

where A and B represent the visible and infrared frames, respectively. The frame size is $M \times N$. Q^{AF} and Q^{BF} are calculated as follows:

$$Q^{AF}(n,m) = Q_g^{AF}(n,m)Q_\alpha^{AF}(n,m) \qquad (21)$$

where $Q_g^{AF}(n,m)$ and $Q_\alpha^{AF}(n,m)$ denote the edge strength and orientation preservation values, respectively, at pixel (n,m). Q^{BF} is defined similarly to Q^{AF}. The weighting factors $\omega^{A}(n,m)$ and $\omega^{B}(n,m)$ reflect the significance of $Q^{AF}(n,m)$ and $Q^{BF}(n,m)$, respectively. It should be noted that $Q^{AB/F}$ is normalized to the range [0, 1] and that a better-quality image has a larger $Q^{AB/F}$ value.

(3) The mutual information is defined as follows:

$$MI_{AF} = \sum_{a,f} p_{AF}(a,f)\log\frac{p_{AF}(a,f)}{p_A(a)p_F(f)} \qquad (22)$$

where A and F represent one of the source images and the fusion result, respectively; p_{AF} is the jointly normalized histogram of A and F; p_A and p_F are the normalized histograms of A and F, respectively; and a and f represent the pixel values in images A and F, respectively. A larger MI indicates a higher correlation between the source image and the fused image.

(4) The structural-similarity-based metric is defined as follows:

$$SSIM_{AF} = \frac{(2\mu_A\mu_F + C_1)(2\sigma_{AF} + C_2)}{(\mu_A^2 + \mu_F^2 + C_1)(\sigma_A^2 + \sigma_F^2 + C_2)} \qquad (23)$$

where A and F represent one of the source images and the fusion result, respectively; μ_A and μ_F represent the mean intensities of images A and F, respectively; σ_A and σ_F are the standard deviations of images A and F, respectively; and C_1 and C_2 are two parameters that are used to avoid instability, both of which are equal to $(KL)^2$, where L is the range of pixel values and $K << 1$ is a small constant. The SSIM represents the structural similarity between the two images, and a larger SSIM value indicates a greater similarity.

Table I QUANTITATIVE ASSESSMENT OF DIFFERENT SOURCE FRAMES.

Source Database	Index	visible frame	infrared frame
(a)OTCBVS	SD	**64.24**	62.31
	EN	**7.73**	6.78
(b)TNO	SD	**45.73**	29.91
	EN	**7.47**	6.53
(c)TNO	SD	48.86	**68.80**
	EN	5.83	**6.59**
(d)INO	SD	42.71	**54.12**
	EN	5.55	**6.05**
(e) INO	SD	51.27	**58.39**
	EN	5.06	**6.31**
(f)Eden	SD	**17.76**	17.56
	EN	**6.02**	5.76

.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2711422, IEEE Transactions on Multimedia

TABLE II QUANTITATIVE ASSESSMENT OF DIFFERENT IMAGE FUSION METHODS. THE SUBSCRIPTS V/F AND I/F REPRESENT THE VISIBLE/FUSED FRAMES AND THE INFRARED/FUSED FRAMES, RESPECTIVELY. THE NUMBERS IN PARENTHESES DENOTE THE RANK IN TERMS OF THE CORRESPONDING INDEX AMONG ALL ALGORITHMS.

Source Database	Index	LAP	DWT	PCA	GRW	GFF	Our Algorithm
(a) OTCBVS	SD	70.03	57.61	20.12	49.66	58.17	**71.82(1)**
	$Q^{AB/F}$	0.5672	0.53	0.44	0.44	0.74	**0.89(1)**
	$MI_{V/F}/MI_{I/F}$	1.85/1.00	1.29/1.13	2.18/**1.26**	1.46/1.22	1.91/0.95	**3.13(1)**/0.97(5)
	$SSIM_{V/F}/SSIM_{I/F}$	0.63/0.55	0.85/0.60	0.69/0.59	0.59/**0.63**	0.75/0.48	**0.88(1)**/0.38(6)
(b) TNO	SD	45.59	25.78	**53.35**	27.57	45.34	47.51(2)
	$Q^{AB/F}$	0.61	0.29	0.64	0.32	**0.71**	0.66(2)
	$MI_{V/F}/MI_{I/F}$	3.92/0.12	1.61/0.27	3.75/0.38	1.42/0.15	**6.83**/0.13	4.59(2)/0.13(4)
	$SSIM_{V/F}/SSIM_{I/F}$	0.94/0.33	0.77/0.61	0.98/0.17	0.80/0.51	0.98/0.26	**0.99(1)**/0.33(3)
(c) TNO	SD	29.90	34.78	65.75	32.06	51.87	**68.69(1)**
	$Q^{AB/F}$	0.50	0.49	0.48	0.28	0.53	**0.62(1)**
	$MI_{V/F}/MI_{I/F}$	0.18/2.26	1.28/0.74	0.26/3.40	0.73/0.52	**3.05**/0.22	0.70(4)/**3.73(1)**
	$SSIM_{V/F}/SSIM_{I/F}$	0.54/0.73	0.80/0.57	**0.95**/0.31	0.56/**0.78**	0.85/0.47	0.78(4)/0.75(2)
(d) INO	SD	51.75	26.92	60.87	25.97	30.31	**62.18(1)**
	$Q^{AB/F}$	0.52	0.51	0.21	0.32	**0.57**	0.54(2)
	$MI_{V/F}/MI_{I/F}$	2.18/1.60	1.65/1.93	0.90/0.75	1.37/1.65	1.57/1.63	**1.48(4)/2.29(1)**
	$SSIM_{V/F}/SSIM_{I/F}$	**0.86**/0.33	0.51/0.72	0.11/0.01	0.42/0.84	0.49/0.75	0.44(3)/**0.91(1)**
(e) INO	SD	54.79	52.73	48.19	54.38	60.09	**67.34(1)**
	$Q^{AB/F}$	0.69	0.61	0.59	0.64	0.64	**0.71(1)**
	$MI_{V/F}/MI_{I/F}$	1.35/2.67	1.19/2.46	2.07/2.49	**2.13**/2.18	1.57/**3.05**	1.30(5)/2.87(2)
	$SSIM_{V/F}/SSIM_{I/F}$	**0.71**/0.71	0.52/0.84	0.61/0.86	0.64/0.73	0.62/0.87	0.58(5)/**0.90(1)**
(f) Eden	SD	15.41	13.50	13.70	13.27	**17.22**	16.83(2)
	$Q^{AB/F}$	0.50	0.44	0.49	0.35	**0.61**	0.50(2)
	$MI_{V/F}/MI_{I/F}$	0.67/0.52	0.62/0.48	0.68/0.50	0.40/**0.93**	0.16/2.46	**1.03(1)**/0.52(3)
	$SSIM_{V/F}/SSIM_{I/F}$	0.67/0.82	0.66/0.84	0.68/0.84	0.61/0.89	0.49/**0.97**	**0.82(1)**/0.84(3)

TABLE III AVERAGE OBJECTIVE PERFORMANCE.

Source Database	Index	LAP	DWT	PCA	GRW	GFF	Our Algorithm
(a)OTCBVS	SD	65.31	63.32	40.23	58.82	69.30	**73.58(1)**
	$Q^{AB/F}$	0.78	0.52	0.67	0.74	0.80	**0.82(1)**
(b) INO	SD	51.39	41.18	47.61	38.56	56.24	**58.51(1)**
	$Q^{AB/F}$	0.63	0.43	0.62	0.48	0.71	**0.77(1)**
(c) Eden	SD	58.34	44.14	55.32	49.39	60.72	**68.69(1)**
	$Q^{AB/F}$	0.64	0.59	0.51	0.58	0.79	**0.83(1)**

. Table I presents the objective performance measures for the original frames. Based on Figs. 4 – 9, it is obvious that in datasets (a), (b) and (f), the visible frames contain more information, whereas in datasets (c), (d) and (e), the infrared frames contain more information. The reason why the visible frames are of relatively poor quality in (c), (d) and (e) is that (c) was shot during the evening with a lack of light, (d) contains a lot of black cars, whereas (e) captures a smoky scene. Although (f) was also shot during the evening, the vegetation was close to the camera and thus is clearly visible.

In Table II, the results of all algorithms are evaluated using the five metrics introduced above. The first metric, SD, reflects the image quality of the fused frame. It is obvious that the proposed algorithm performs stably and ranks highly among all competitors in terms of SD. The metric $Q^{AB/F}$ represents the structural information of the fused frame compared with that of the original frames, A and B. The proposed algorithm also performs highly and stably with respect to this metric.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2711422, IEEE Transactions on Multimedia

However, the objective results for the proposed algorithm are unstable in terms of the MI and SSIM metrics. MI and SSIM represent the mutual information and structural similarity between a source frame and the fused frame. From Table II, it is evident that the performance of the proposed algorithm on databases (a), (b) and (f) is relatively high in terms of $MI_{V/F}$ and $SSIM_{V/F}$, whereas on databases (c), (d) and (e), it is relatively high in terms of $MI_{I/F}$ and $SSIM_{I/F}$. Based on Table I, it can be concluded that when the original frame is of better quality than the infrared frame, the fusion results obtained using the proposed algorithm will have better $MI_{V/F}$ and $SSIM_{V/F}$ values, whereas when the original infrared frame is of better quality than the visible frame, the fusion results will have better $MI_{I/F}$ and $SSIM_{I/F}$ values. Although the proposed algorithm does not always show the best performance in terms of the MI and SSIM metrics, it can adaptively extract the most useful information based on the video quality.

To evaluate the proposed algorithm more objectively, experiments were conducted on 50 pairs of frames chosen from three of the datasets, i.e., 20 pairs from OTCBVS, 15 pairs from INO and 15 pairs from Eden. It should be noted that OTCBVS has 1054 frames in total, INO has 2048 frames in total, Eden has 1700 frames in total. The average objective performance is summarized in Table III. As is shown in Table III, the proposed algorithm has an obvious advantage over the other algorithms. It should be noted that the MI and SSIM metrics are not included in Table III because the purpose of these metrics is to show the ability of the fusion result to adapt to the source images, and they are meaningless for evaluating the average performance.

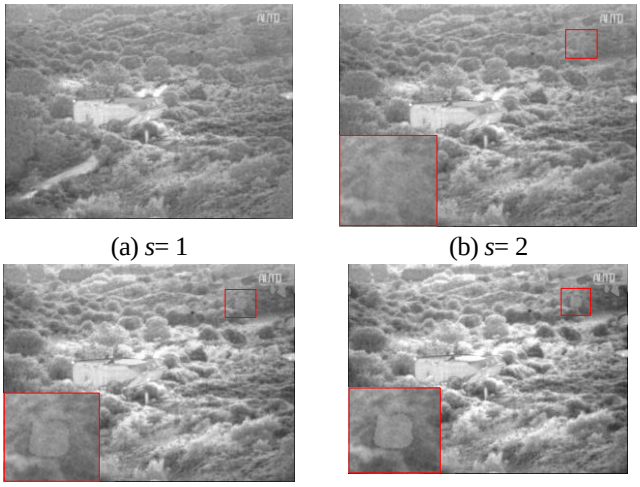
5. Discussion

A. Discussion of the parameters of the proposed algorithm

In this section, both the experimental results and the objective evaluation results for different values of the parameters of the proposed algorithm will be presented. There are two important parameters in the proposed algorithm, namely, *s* in equation (13) and *d* in equation (18). It should be noted that for the analysis of the influence of *s*, the value of *d* was set to 1, whereas for the analysis of the influence of *d*, the value of *s* was set to 2. The parameter *s* is used to control the tradeoff between artifacts and hot objects. As mentioned before, if *s* is too large, then hole artifacts will be introduced. If *s* is too small, regions that are not sufficiently hot will affect the fusion. A comparison of results obtained with different values of *s* is shown in Fig. 12. When *s* is set to 1, the generated result has poor illumination conditions, and the visible textures are not clear, which indicates that the result is affected by infrared regions that are not sufficiently hot. When *s* is set to an excessively large value, white spots appear in the results, which means that hole artifacts have been introduced. These white spots can be clearly seen in the close-up views shown in the bottom left corners of the results. Therefore, 2 is an appropriate default value for *s*

The parameter *d* is used to control the extent to which visible textures are added into the base layer fusion result. The choice of *d* depends on the quality of the visible video. If the visible video needs detail enhancement, then a higher *d* value can be chosen. However, a higher *d* value will sometimes result in increased noise. A comparison of results obtained with different values of *d* is shown in Fig. 13. When *d* is set to 0.5, the textures are weakened.

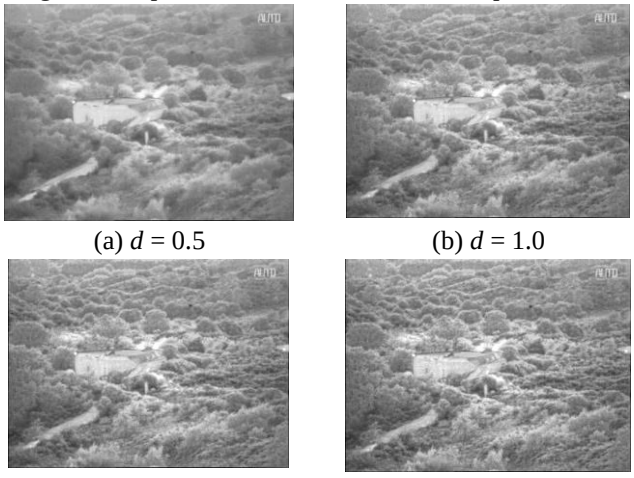
The assignment of a value of 1 for *d* means the result contains the same textures as the original visible frame, which is a suitable result. When *d* is set to 1.3, it can be seen that the visible textures are enhanced. However, when *d* is set to 1.5, the result seems to be over-enhanced, and the noise is enhanced as well. According to Fig. 13, the value of the parameter *d* should be cautiously chosen, and a small change in *d* will lead to a large difference in the result.



(a) *s*= 1 (b) *s*= 2

(c) *s*= 10 (d) *s*= 100

Fig. 12 Comparison of results with different parameter *s*.



(a) *d* = 0.5 (b) *d* = 1.0

(c) *d* = 1.3 (d) *d* = 1.5

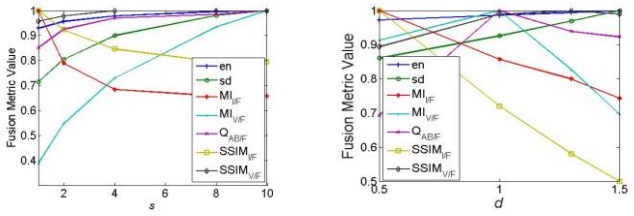
Fig. 13 Comparison of results with different parameter *d*.

To objectively compare the influences of different parameter values, the objective fusion quality metrics were calculated based on Fig. 12 and Fig. 13. All values were normalized to clearly show the influences of the different parameter values. Fig. 14 presents the objective evaluation results. As shown in Fig. 14(a), a larger s value results in a higher similarity between the visible and fused frames and a smaller similarity between the infrared and fused frames,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2711422, IEEE Transactions on Multimedia

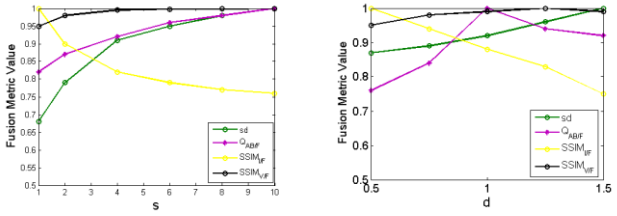
reflecting the same performance observed in the experiments. By contrast, both the SD and EN metrics increase as s grows larger because the SD and EN metrics for the visible frame are larger than those for the infrared frame, and when s increases, the fusion result becomes more similar to the visible frame. However, the results will contain severe artifacts when s is too large. From Fig. 14(b), it can be observed that some metrics, such as SSIMV/F, MIV/F and QAB/F, have their maximum values when d = 1; this is because both the weakening of the textures and the enhancement of the textures will affect the similarity between the visible frame and the fused frame.

Moreover, to verify the robustness of the metrics curve variation in Fig. 14, we extend the test set from single frame to multiple diverse frames in OTCBVS and TNO dataset. As shown in Fig.15(a)-(b), for SSIMV/F, SD and QAB/F, the metrics variation trend is consistent with the result in Fig.14(a)-(b)



(a) s (b) *d*

Fig. 14 Performance of the proposed algorithm with different algorithms for figure 13 and 14. i.e., (a) s, (b) d.



(a) *s* (b)d

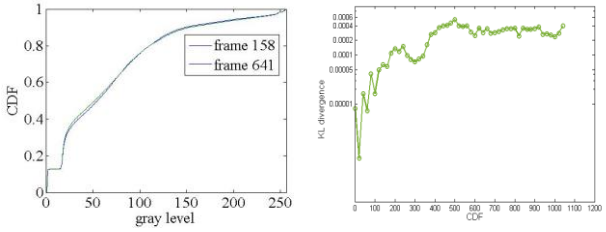
Fig. 15 Performance of the proposed algorithm with different algorithms for multiple diverse frames. i.e., (a) *s*, (b) *d*.

B. An extension of the proposed algorithm

Although the proposed algorithm yields an appropriate fusion result with a relatively low computational complexity, it is far from satisfying the requirements for real-time applications. Videos show inter-frame correlations, and both the entropy and the CDF of the gray levels may be similar
among adjacent frames. Thus, a real-time video fusion algorithm is proposed that utilizes the inter-frame correlations.

In single-frame fusion, the CDF is generated to represent the global information of the frame. Meanwhile, in a video without any scene changes, such as a surveillance video, the global information among the adjacent frames will remain stable. And the local area variation mainly from objects will not impact the global information, which can be confirmed by Fig.16. As shown in Fig.16(d), the KL divergence between the first frame and the subsequence frames are small and fluctuate in a relatively small range. Therefore, the CDFs for different frames representing the same scene are regarded as identical in this paper. Consequently, the CDF is calculated only once, for the first frame in a given scene. Any change of scene, which can be detected based on sensor parameters in a video surveillance application, will necessitate a re-calculation of the CDF.

(a) Frame 158 (b) Frame 641



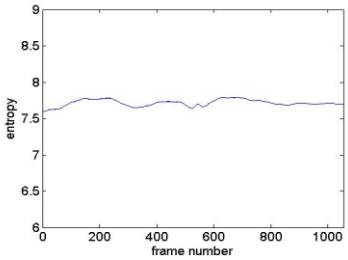
(c) CDF (d)CDFs KL

Fig. 16 CDF comparison between different frames.

In single-frame fusion, the entropy is generated for the adaptive adjustment of the visible weight map. Meanwhile, in a video, the entropy of a frame will predominantly depend on the sensor technology, the illumination conditions, and so on. Moving hot objects will have almost no effect on the entropy, so the entropy in a video should be stable. Fig. 17 shows an example of the entropy in the individual frames of a video with 1054 frames. From Fig. 17, it can be concluded that the entropy remains reasonably stable throughout a video; therefore, the entropy can also be calculated only once, for the first frame in a given scene. Again, a scene change will necessitate a re-calculation of the entropy.

Moreover, the guided filter has a time complexity of *O(N)* with respect to the number of pixels N, independent of the filter size. He et al [31] propose that by means of subsampling, a guided filter can be sped up to run in $O(N/s^2)$ time.

(a) Frame 1 (b) Frame 537

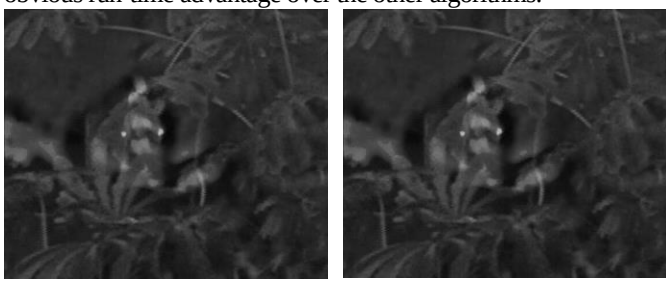


(c) Entropy

Fig. 17 The entropy of the visible video.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2711422, IEEE Transactions on Multimedia

For an evaluation of the run time of the proposed algorithm, Table IV shows the average run times on three datasets. The experiments were performed on a computer with a 2.67 GHz CPU and 4 GB of memory. In Table IV, the run times for the LAP, GRW and GFF algorithms are compared with those of the proposed algorithm based on the experiments described above. "Our algorithm (frame)" indicates the run time of the proposed single-frame fusion algorithm, whereas "Our algorithm (video)" indicates the run time of the proposed video fusion algorithm using inter-frame correlations and the fast guided filter. From Table IV, it can be concluded that the proposed algorithm has an obvious run-time advantage over the other algorithms.



(a) Our algorithm (frame) (b) Our algorithm (video)

Fig. 18 Comparison of results between the original single-frame fusion algorithm and the sped-up video fusion algorithm.

Table IV Run times of different algorithms (seconds).

	OTCBVS	INO	Eden
LAP	0.04	0.11	0.08
DWT	0.03	0.22	0.35
PCA	0.12	0.13	0.18
GRW	0.09	0.12	1.53
GFF	0.73	1.15	1.13
Our algorithm (frame)	0.12	0.22	0.38
Our algorithm (video)	**0.02**	**0.03**	**0.05**

TABLE V COMPARISON OF THE AVERAGE OBJECTIVE PERFORMANCE BETWEEN THE ORIGINAL SINGLE-FRAME FUSION ALGORITHM AND THE SPED-UP VIDEO FUSION ALGORITHM.

Source Database	Index	Our Algorithm (frame)	Our Algorithm (video)
(a) OTCBVS	SD	73.58	72.00
	$Q^{AB/F}$	0.82	0.83
(b) INO	SD	58.51	56.73
	$Q^{AB/F}$	0.77	0.76
(c) Eden	SD	68.69	68.10
	$Q^{AB/F}$	0.83	0.83

Moreover, experimental results and objective evaluation metrics are presented to compare the results of the original single-frame fusion algorithm and the sped-up video fusion algorithm. The results are shown in Fig. 16 and Table V. Fig. 17 shows that there is almost no visual difference between the results of the original single-frame fusion algorithm and the sped-up video fusion algorithm. Regarding the average objective performance measures, some differences are evident between the two fusion algorithms because of the minor instability in the CDF and information entropy over time. However, this instability has little effect on the experimental results, which are acceptable.

6. Conclusion

This paper proposes an adaptive fusion algorithm for visible and infrared videos. The original visible and infrared frames are decomposed into two layers, namely, the base layer and the detail layer, using a guided filter. The visible base layer and the infrared base layer are fused based on their corresponding weight maps generated using the proposed adaptive approach, and the final fusion result is obtained by combining the fused base layer with the visible detail layer. The proposed algorithm can achieve adaptive fusion such that both infrared-hot objects and visible textures are retained. Experimental results based on four public datasets demonstrate that the proposed fusion algorithm achieves better fusion results compared with state-of-the-art methods.

7. Reference

[1] Liang Xu, Junping Du, and Zhenhong Zhang, "Infrared-visible video fusion based on motion-compensated wavelet transforms," IET Image Processing, 2015.

[2] Rikke Gade and Thomas B. Moeslund: "Thermal cameras and applications: a survey", Machine Vision and Applications, 25:245-262, 2014

[3] Zhang Q, Wang L, and Ma Z, "A novel video fusion framework using surfacelet transform," Optics Communications, 3032-3041,2012.

[4] Zhang Q, Chen Y, and Wang L, "Multisensor video fusion based on spatial–temporal salience detection," Signal Processing, 2485-2499, 2013.

[5] D. P. Bavirisetti and R. Dhuli, "Fusion of Infrared and Visible Sensor Images Based on Anisotropic Diffusion and Karhunen-Loeve Transform", IEEE Sensors Journal, 203 – 209, 2016.

[6] Jiao D, Weisheng Li, Bin X and Qamar N, "Union Laplacian pyramid with multiple features for medical image fusion," Neurocomputing, 326–339, 2016.

[7] TAO Wan, N. Canagarajah and A. Achim, "Segmentation-Driven Image Fusion Based on Alpha-Stable Modeling of Wavelet Coefficients," IEEE Transactions on Multimedia (TMM), 624-633, 2009.

[8] Mahyari A G and Yazdi M, "A novel image fusion method using curvelet transform based on linear dependency test," IEEE International Conference on Digital Image Processing, 2009.

[9] Yang L, Guo B L, and Ni W, "Multimodality medical image fusion based on multiscale geometric analysis of contourlet transform," Neurocomputing, 203-211, 2008.

[10] G. Bhatnagar, Q. M. Jonathan Wu and Zheng Liu, "Directive Contrast Based Multimodal Medical Image Fusion in NSCT Domain," IEEE Transactions on Multimedia (TMM), 1014-1024, 2013.

[11] Shen R, Cheng I, and Shi J, "Generalized random walks for fusion of multi-exposure images," IEEE Transactions on Image Processing(TIP), 3634-3646, 2011.
[12] T. Shibata, M. Tanaka, and M. Okutomi, "Visible andnear-infrared image fusion based on visually salient areaselection," Proc. of IS&T/SPIE Electronic Imaging, p.94040G, 2015.
[13] Ji Wang, S. Acharya and M. Kamt, "Adaptive decision fusion using genetic algorithm," IEEE Annual Conference on Information Science and Systems (CISS), 2016.
[14] Li S, Kang X, and Hu J, "Image fusion with guided filtering," IEEE Transactions on Image Processing (TIP), 2864-2875, 2013.
[15] Yong M, Jun C, Chen C, Fan F and Jiayi M, "Infrared and visible image fusion using total variation model," Neurocomputing, 12-19, 2016.
[16] He K, Sun J, and Tang X, "Guided image filtering," Pattern Analysis and Machine Intelligence (PAMI), 1397-1409, 2013.
[17] Kordelas, Georgios A., et al, "Content-Based Guided Image Filtering, Weighted Semi-Global Optimization, and Efficient Disparity Refinement for Fast and Accurate Disparity Estimation," IEEE Transactions on Multimedia, 155-170, 2016.
[18] Y. Ding, J. Xiao, and J. Yu, "Importance Filtering for Image Retargeting," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 89-96, 2011.
[19] R. Brink, "Using spatial information as an aid to maximum entropy image threshold selection," Pattern Recognition Letters 17 (1996) 29-36.
[20] G. Boccignone et al., "Encoding Vis4ual Information Using Anisotropic Transformations," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Vol. 23, No. 2, 2001.
[21] Shibata T and Tanaka M, "Unified image fusion based on application-adaptive importance measure," IEEE International Conference on Image Processing (ICIP), 2015.
[22] Davis J W and Sharma V, "Background-subtraction using contour-based fusion of thermal and visible imagery," Computer Vision and Image Understanding (CVIU), 162-182, 2007.
[23] Toet A, Hogervorst M A and Lensen H A, "ATHENA: The combination of a brightness amplifier and thermal viewer with color," TNO defense security and safety SOESTERBERG (NETHERLANDS), 2007.
[24] www.ino.ca/en/video-analytics-dataset/
[25] Lewis J J, Nikolov S G, and Loza A, "The Eden Project multi-sensor data set," The Online Resource for Research in Image Fusion (ImageFusion. org), 2006.
[26] BBurt P J and Adelson E H, "The Laplacian pyramid as a compact image code," IEEE Transactions on Communications,532-540,1983.
[27] Rajenda Pandit Desale and Sarita V. Verma, "Study and Analysis of PCA, DCT & DWT based Image Fusion Techniques", Signal Processing Image Processing & Pattern Recognition (ICSIPR), 66-69, 2013.
[28] C. Xydeas and V. Petrovi´ c, "Objective image fusion performance measure," Electron. Lett., vol. 36, no. 4, pp. 308–309, Feb. 2000.
[29] Qu G, Zhang D, and Yan P, "Information measure for performance of image fusion," Electronics letters, 313-315, 2002.
[30] Wang Z, Bovik A C, and Sheikh H R, "Image quality assessment: from error visibility to structural similarity," IEEE Transactions on Image Processing, 600-616, 2004.
[31] He K, Sun J, "Fast guided filter," arXiv:1505.00996, 2015.

8. ***Author***

Hai-Miao Hu received the B.S. degree from Central South University, Changsha, China, in 2005, and the Ph.D. degree from Beihang University, Beijing, China, in 2012, all in computer science. He was a visiting student at University of Washington from 2008 to 2009. Currently, he is an associate professor of Computer Science and Engineering at Beihang University. His research interests include video coding and networking, image/video processing, and video analysis and understanding.

Jiawei Wu received the B.S. degree in Computer Science and Technology from the College of Computer, Xidian University, Xi'an, China, in 2014, and he is currently pursuing the M.S. degree in computer science and engineering from Beihang University, Beijing, China. His current research interests include video fusion and video enhancement.

Bo Li received the B.S. degree in computer science from Chongqing University in 1986, the M.S. degree in computer science from Xi'an Jiaotong University in 1989, and the Ph.D. degree in computer science from Beihang University in 1993.Now he is a professor of Computer Science and Engineering at Beihang University, the Director of Beijing Key Laboratory of Digital Media, and has published over 100 conference and journal papers in diversified research fields including digital video and image compression, video analysis and understanding, remote sensing image fusion and embedded digital image processor.

Guo Qiang, received the B.E. degree in computer science and technology from Sichuan University, Chengdu, China, in 2016. He is currently working towards the M.S. degree in computer science and technology from Beihang University, Beijing, China. His researches interests include computer vision, image enhancement, image processing, and intelligent video analyzing system.

Jin Zheng was born in Sichuan, China, on October 15, 1978. She received the B.S. degree in applied mathematics and informatics from the College of Science in 2001, and the M.S. degree from the School of Computer Science, Liaoning Technical University, Fuxing, China, in 2004, and the Ph.D. degree from the School of Computer Science and Engineering, Beihang University, Beijing, China, in 2009.She is currently a Teacher with Beihang University. Her current research interests include moving object detection and tracking, object recognition, image enhancement, video stabilization, and video mosaic.