Journal of Computers Vol. 32 No. 6, 2021, pp. 52-65
doi:10.53106/199115992021123206005

Infrared And Visible Image Fusion Based on Rolling Guidance Filter Combined with Convolutional Neural Network

Jin-Peng Dai[1,2], Zhong-Qiang Luo[1,2*], Cheng-Jie Li[3]

[1] School of Automation and Information Engineering, Sichuan University of Science and Engineering, Yibin 644000, PRC
1078454292@qq.com

[2] Artificial Intelligence Key Laboratory of Sichuan University of Science and Engineering, Yibin 644000, PRC

[3] School of Computer Science and Technology, Southwest Minzu University, Chengdu 610041, PRC

Received 9 March 2021; Revised 15 July 2021; Accepted 15 August 2021

Abstract. Infrared and visible image fusion has been a hot issue in the field of image processing. By fusing the infrared image and visible image of a same scene, the thermal radiation information of the infrared image and the texture information of the visible image can be retained in one image. However, the image fusion of infrared and visible images will lead to unclear image texture and loss of deep details. The conventional multi-scale transform methods will lead to the fuzzy edge detail of fused image. To solve these problems, an image fusion algorithm based on edge preserving filter combined with convolutional neural network is proposed. First, the source images are decomposed into two-scale images by rolling guidance filter. Second, data normalization is used for generating weights of base layers fusion. Third, the activity weight maps of detail layers are generated by VGGnet. Finally, different fusion strategies with different scales are used for image reconstruction. Compared with the existing representative methods, the proposed model performs well in both subjective and objective evaluation, especially in information entropy and edge detail preservation.

Keywords: image fusion, rolling guidance filter, VGGnet, image reconstruction

1 Introduction

Infrared and visible image fusion refers to the fusion of infrared and visible images of the same scene into an image, which is a multi-modal data fusion [1]. In general, it's hard to get useful information only through visible light images under bad lighting conditions or camouflage. Infrared images come from infrared sensors, which can capture electromagnetic waves of different frequencies radiated by objects, also known as thermal radiation. At the same time, the infrared image will also lose the texture details, because the heat emitted by the object is hardly affected by the texture. Therefore, infrared and visible image fusion simultaneously retains the thermal radiation information and texture information of source image. It is widely used in the field of object detection, object tracking, night color version, biological recognition [2].

Broadly speaking, the current fusion algorithms of infrared and visible image can be divided into five categories: multi-scale transform [3-4], sparse representation [5-6], subspace [7], saliency [8], hybrid method [9], and deep learning method [10-12]. In the past decades, multiscale transform is the most commonly used method in image fusion. The original image is decomposed into different scale components by multi-scale transform. The multi-scale components of the source image is fused according to the given fusion rules, and then the corresponding multi-scale inverse transform is used to

* Corresponding Author

construct the fused image. The multi-scale transform method has the same visual characteristics as human eyes system, this makes the fused image have better visual effect. The common multi-scale transform methods include pyramid transform, wavelets transform, non-subsampled contourlet transform and curvelet transform etc. Sparse representation uses a super complete dictionary to represent the source image sparsely. At the same time, the sliding window strategy is used to divide the source image into multiple overlapping blocks, which potentially enhances the stable image representation and improves the robustness of fusion. Zhu et al. [6] proposed a fusion framework that integrated image-patches clustering and online dictionary learning in sparse representation process. This method optimizes the dictionary by clustering to improve the fusion quality. The subspace-based method is to project the source image from high dimension to low dimension subspace. For most source images, there are lots of redundant information. Projection into low dimensional space can help to observe the internal structure. The saliency-based method stems from a well-established fact: if the pixel gray value of a region is greatly different from the adjacent regions, it can be considered that the visual effect of this region is more significant from the overall effect. The saliency-based method helps to maintain the integrity of salient regions and improve the visual quality of the fused image. The hybrid method combines the advantages of two methods. Huang et al. [9] proposed a fusion method based on non-subsampled contourlet transform (NSCT) and neural network. In their framework, NSCT is used to decompose the source image into low frequency and high frequency. Maximum absolute value and pulse coupled neural network (PCNN) fusion rules are applied to high frequency component fusion, low frequency component fusion is designed to enhance the detailed features of the fused image according to the activity measures. The fused image is obtained by inverse transform. Huang's method achieves a good fusion effect, it can still be optimized from algorithm efficiency and weight settings.

In recent years, with the development of computer vision field, deep learning began to be used in image fusion. In ICCV 2017, Prabhakar et al. [13] proposed deepfuse to solve the problem of multi exposure fusion. They designed an encoder with two convolution layers. The features of the source image were added separately, and then the fusion results were output through the decoding layer of three convolution layers. This method achieved good results in multimodal image fusion. However, because of the simple network model, the deep information of the source image is not effectively utilized. Li et al. [14] proposed a novel deep learning architecture based on CNN layers and dense block. In his model, the feature maps which are obtained by each layer in encoding network are cascaded as the input of the next layer. In addition, Li et al. [11] proposed a fusion method of infrared and visible light based on deep learning framework. In his method, the source image is divided into the base layer and the detail layer. The weighted average method is used to the basic level image fusion, and the VGG-19 network is used to extract the depth features in the detail layer to generate the image with high quality. However, although this method has good robustness, it reduces the contrast of the image. To solve these problems, Li et al. [15] proposed a novel fusion framework based on residual network (ResNet) and zero-phase component analysis (ZCA) to fully utilize and process the deep features. The deep feature output of ResNet already contains multi-layer information, so this method only uses single layer output. Then the deep features are projected to the sparse domain by ZCA operation, and the weight map is obtained by l_1-norm operation. The fusion image is reconstructed by the final weight map and the source image. Gao et al. [16] used transfer learning to change the network structure of deep learning framework. It makes VGG-19 network extract features from infrared images more accurately. Recently, an improved fusion method was proposed by Yan et al. [17]. In order to get better image fusion based on saliency, the image fusion is based on saliency. Multi-resolution singular value decomposition (MVD) is performed in detail layer, this makes the reconstructed image keep the details of the source image effectively. Although these methods have achieved good fusion results, the image texture and details can not be preserved well, and the setting of fusion strategy will lead to the loss of energy.

To overcome these drawbacks, a novel model based on rolling guidance filter [18] (RGF) and CNN layers is proposed. This is the first time that RGF and deep learning methods have been combined in the field of image fusion. Firstly, we get the base layer and the detail layer of the infrared and visible image pairs through RGF. In our network, detail layer image pairs are used as inputs. Then, due to the different contents of the basic layer and the detail layer, we design different fusion strategies to maintain the characteristics of different layers of images. The key contributions of this paper can be elaborated as follows:

- Rolling guidance filter is used to decompose the source images into different layers. This filtering operation can keep the edge information of the source images.
- A new weighted fusion strategy for base layer is proposed, which causes the texture information of the fused image more reasonably.
- VGGnet is applied for detail extraction. By processing the outputs of convolution blocks, the weight map of detail layer fusion is obtained.

In Section 2, we review related work while Section 3 we describe our proposed algorithm. The experimental results and analysis are shown in Section 4. Finally, Section 5 draws the conclusions to the paper.

2 Related Work

2.1 Rolling Guidance Filter

Infrared and visible images contain many levels of important structures and edges. Therefore, it is very important to select an method to smooth the image while maintaining different hierarchical structure. In the field of image processing and computer vision, for instance, bilateral filter, guided filter, weighted least squares filter, geodesic filter and so on is the mainstream of edge-preserving filters [19]. Although the above methods have been applied in the field of image fusion, and achieved better results. There are still some hard problems to be solved: edge-preserving methods hardly separate structure from details, because edge strength and object scale are completely different concepts.

Rolling guidance filter can smooth small structures and restore edges, It is based on a rolling guidance implemented in an iterative manner that converges quickly [18]. This makes it more efficient than other filters of the same type.

The first step is to remove small structures (size less than σ_S) of X_k by Gaussian filter, and obtain the original guidance image G_k, where $k \in \{I, V\}$, I is the infrared images, and V is the visible images. The filter of pixel p in image X_k can be represented as follows:

$$G_k(p) = \frac{1}{U_p} \sum_{q \in N(p)} exp(-\frac{\| p-q \|^2}{2\sigma_S^2}) X_k(q), \tag{1}$$

where

$$U_p = \sum_{q \in N(p)} exp(-\frac{\| p-q \|^2}{2\sigma_S^2}), \tag{2}$$

is for normalization. N_p is the set of neighboring pixels of p, and q belongs to N_p. σ_S is the structure scale parameter, which is defined as the smallest standard deviation of Gaussian kernel. The second step is iterative edge recovery. In this step, image J is iteratively updated. J^{i+1} is denoted as the result in the i-th iteration. J^1 is set as the output of Gaussian filter, which is represented as G_k in Eq. (1).

$$J^{i+1}(p) = \frac{1}{K_p} \sum_{q \in N(p)} exp(-\frac{\|p-q\|^2}{2\sigma_S^2} - \frac{\|J^i(p)-J^i(q)\|^2}{2\sigma_R^2}) X_k(q), \tag{3}$$

where

$$K_p = \sum_{q \in N(p)} exp(-\frac{\| p-q \|^2}{2\sigma_S^2} - \frac{\| J^i(p)-J^i(q) \|^2}{2\sigma_R^2}), \tag{4}$$

is used for normalization. σ_R is the range weights parameter. This process essentially uses bilateral filtering to iteratively change the guidance image. This operation is named as rolling guidance. Since this process uses J^i to compute the affinity between pixels, it makes resulting structures similar to J^i. It yields structure transform from J to X_k.

2.2 Convolutional Layer of VGG-19

The VGG network [20] is a deep convolution neural network developed by Oxford Computer Vision Group and Google deepmind. VGG-19 network consists of five convolution layer blocks, followed by three full connection layers. The convolution layer is mainly responsible for extracting the feature map of the image, and the whole connection layer is used as the classification function.

For each convolutional layer, the number of channels used are different. The Conv1 block contains two convolutional layers (each with 64 channels). The Conv2 block also contains two convolutional layers, but the number of channels layer has increased to 128. The Conv3 block contains four convolutional layers (each with 256 channels). And the Conv4 block contains four convolutional layers with 512 channels. In our paper, only the first four layers of VGG-19 network are used in our model. The rest of the network structure will not be described in detail.

3 Proposed Method

The proposed method concludes four steps, which are presented at Fig. 1. The proposed method consists of four stages, each of which will be explained in detail in the following sections.

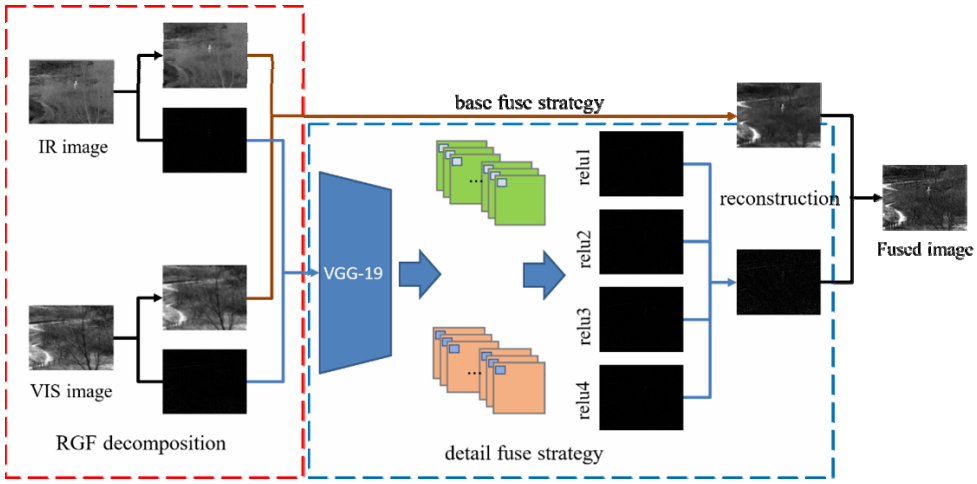


Fig. 1. Schematics of the proposed IR and VIS images fusion framework

3.1 Image Decomposition

The source images are represented as $\boldsymbol{X_k}$, where $k \in \{I, V\}$, I and V are refer to infrared and visible images repectively. After the iteration of rolling guidance filter,The base images B_k can be obtained directly by multi-scale transformation. This process is described in Fig. 2.

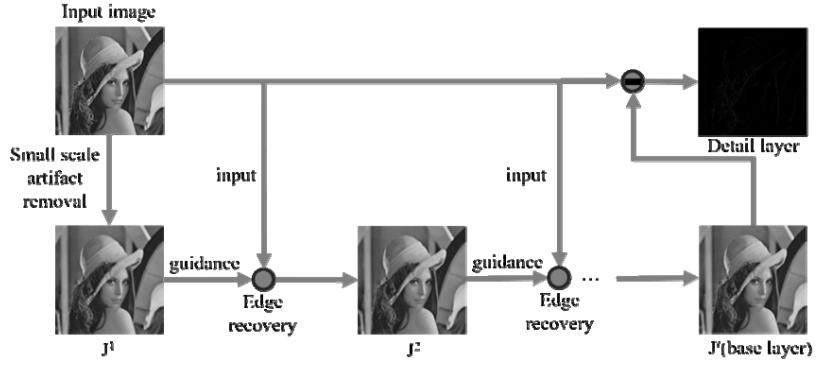


Fig. 2. Flow chart of rolling guidance filter decomposition

Image decomposition is expressed by formula:

$$B_k = RGF(X_k, \sigma_S, \sigma_R, i), \tag{5}$$

where σ_S and σ_R respectively expressed as structure scale and range weight parameter. i is the number of the iterations. The detail layer can be obtained by the following equation:

$$D_k = X_k - B_k, \tag{6}$$

where Dk is the detail image. Therefore, the base layer and detail layer of source images can be get.

3.2 Base Layer Fusion

The design of reasonable fusion rules is the key technology of image fusion algorithm. It determines the selection and fusion of the final image pixel or coefficient, which is very essential to the fusion effect. In this section, we propose a method to normalize the base layer images and get the weights for base layer fusion.

Firstly, the base images B_k are transformed from the two-dimensional matrix to the single line matrix B'_k. $m \times n$ is the size of base images. Set the x-th ($x \in \{1, 2, ..., m\}$) elements of B_k as the $(((x-1)\times n+1):(x\times n))$ segment elements of B'_k. The size of B'_k is $1\times(m\times n)$. The *mapminmax*(·) function is used for normalizing the matrix B'_k. The normalized results are the weight of each pixel in the corresponding base images.

$$W_{B_k}(x,y) = mapminmax(B_k', \alpha, \beta) = \frac{B_k'(x,y) - min(B_k')}{max(B_k') - min(B_k')}. \tag{7}$$

In the above equation, W_{Bk} is the weight of base layer images, α, β is parameters to control the normalized range. α, β are separately set to 0 and 1, which means the data is normalized to intervals 0 to 1 by the given rule. $max(B'_k)$ and $min(B'_k)$ respresent the maximum and minimum of B'_k respectively. The following equation represents the base layer fusion strategy.

$$F_B(x,y) = \sum_{k\in\{I,V\}} W_{B_k}(x+(y-1)\times n)\times B_k(x,y). \tag{8}$$

3.3 Detail Layer Fusion

The fusion strategy of detail layer is to obtain the activity level map of detail level images through VGG-19 network. The diagram is shown in the Fig. 3.

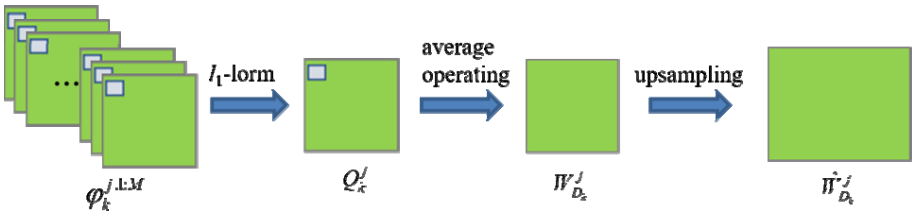


Fig. 3. The diagram of the extraction process for the detail layer weight map

As mentioned above, VGG-19 network has four convolution layers. According to the formula:

$$\varphi_k^{j,1:M} = \Phi_j(D_k), \tag{9}$$

where $\varphi_k^{j,1:M}$ represents the M-th channels map of input images in j-th convolution layer, $j \in \{1,2,3,4\}$, $M=64\times 2^{j-1}$, Φ_j refers to the j-th layer of VGG-19. The same action is used for each layer. l_1-norm is used to measure the activity level of detail content. Therefore, the initial activity level map Q_k^j can be obtained by the following operations.

$$Q_k^j(x,y) = \| \varphi_k^{j,1:M}(x,y) \|_1 . \tag{10}$$

Then, the block based averaging operator is used to calculate the final activity level map $\hat{Q}_k^j$, which makes the results robust.

$$\hat{Q}_k^j(x,y) = \frac{\sum_{\lambda=-\omega}^{\omega} \sum_{\mu=-\omega}^{\omega} Q_k^j(x+\lambda, y+\mu)}{(2\omega+1)^2}, \tag{11}$$

where ω is a parameter determines the block size. The feature weight map $W_{D_k}^j$ can be directly obtained from the final activity level map:

$$W_{D_k}^j = \frac{1}{\sum_{k\in\{I,V\}} \hat{Q}_k^j} \hat{Q}_k^j. \tag{12}$$

Due to the different layers of the network, the size of activity level map is different. An upsampling operation is needed to match the size of the feature weight map with the input detail image. It is calculated by the following equation:

$$\hat{W}_{D_k}^j(x+a, y+b) = W_{D_k}^j(x,y), \tag{13}$$

where $a, b \in \{0, 1, ..., 2^{j-1}-1\}$, which represents the size difference of the image after upsampling. $\hat{W}_{D_k}^j$ is the weight map of detail layel after overall registration. The fusion result of detail layel images is obtained by the following equation:

$$F_D^j(x,y) = \sum_{k\in\{I,V\}} \hat{W}_{D_k}^j(x,y) \times D_k(x,y). \tag{14}$$

3.4 Image Reconstruction

After obtaining the fusion results of base images F_B and detail images F_D^j, using additive strategy for fusing image reconstruction, as shown in Eq. 15.

$$F = F_B + \sum_{j\in\{1,2,3,4\}} F_D^j. \tag{15}$$

4 Experiment and Result Analysis

4.1 Experimental Conditions and Setting

The algorithm experimental environment is as follows: The host is configured with Intel (R) Core (TM) i5. The main frequency is 2.9 GHz and the memory is 16 GB. The experimental simulation platform is MATLAB R2018b. In order to evaluate the performance of the fusion algorithm objectively, We selected several representative infrared and visible image pairs from the most widely used benchmark dataset TNO [21] to conduct experiments. Five representative image pairs from the TNO dataset are presented in Fig. 4

To verify the accuracy of the algorithm in this paper, the algorithm proposed in this paper is compared with the traditional classical algorithm and recently proposed advanced algorithm. The contrast algorithm are as follows: deep learning framework (CNN) [11], rolling guidance filter (RGFF) [19], MDLatent low-rank representation (MDLatLRR) [22], visual saliency map and weight least square (VSMWLS) [23], non-subsampled contourlet transform (NSCT) [24], infrared feature extraction and visual information preservation (IFEVIP) [25], and ResNet fusion (ResFuse) [15]. RGFF and NSCT are

Fig. 4. Source images from the TNO dataset used in our experiments, first row contains visible images while second row contains infrared images

methods of multi-scale transform, MDLatLRR belongs to representation learning, VSMWLS and IFEVIP are saliency based methods, CNN and ResFuse are deep learning methods. All these techniques could attain the appropriate fused images. The comparison of the proposed method with these techniques is therefore relevant, enabling the proposed method to validate its benefits.

4.2 Evaluation of Algorithm

The quality of image fusion is an important index. Due to different application purposes or scenes, fused images have different measurement standards. Therefore, the comprehensive use of a variety of evaluation criteria can better determine the fusion effect. Generally, the fusion quality evaluation methods of infrared and visible image fusion include subjective evaluation and objective evaluation [26].
Subjective Evaluation. Subjective evaluation method is based on human visual system to evaluate the quality of fusion image, which plays an important role in fusion quality evaluation. The subjective evaluation method is to score the fusion image by trained observers, and make artificial evaluation according to the image details, object integrity, image distortion and other standards. This method has the disadvantages of manual intervention, time-consuming, high cost and non-reproducible.
Objective Evaluation. The objective evaluation method can quantitatively and automatically measure the quality of the fusion image, which is more popular and applicable in the scientific research field. Objective evaluation methods come in different types, which are based on information theory, structure similarity, image gradient, statistics, and human visual system. In this paper, five commonly used objective evaluation indexes of infrared and visible image fusion are selected, which are entropy (EN), spatial frequency (SF), standard deviation (SD), average gradient (AG) and mutual information (MI) [2, 27]. These metrics are positive indicators, and the larger the value is, the better the result.
Entropy (EN). The entropy of an image is used to measure the amount of information contained in the image. EN is mathematically defined as follows:

$$EN = -\sum_{l=0}^{L-1} p_l \log_2 p_l, \tag{16}$$

where l is the total gray level of the image, and p_l is the probability density of the first level gray value of the image, which can be calculated from the gray histogram of the image.
Spatial Frequency (SF). Spatial frequency represents the activity of the image and reflects the clarity of the image. The formula is as follows:

$$SF = \sqrt{RF^2 + CF^2}, \tag{17}$$

where RF is the row frequency of the image , and CF is the column frequency of the image. The specific calculation formula is as follows:

$$RF = \sqrt{\frac{1}{m \times n} \sum_{x=1}^{m} \sum_{y=1}^{n} (F(x,y) - F(x,y-1))^2}, \tag{18}$$

$$CF = \sqrt{\frac{1}{m \times n} \sum_{x=1}^{m} \sum_{y=1}^{n} (F(x,y) - F(x-1,y))^2}, \tag{19}$$

where $F(x, y)$ is the pixel value at point (x, y), and the image size is $m \times n$.

Standard Deviation (SD). The standard deviation is reflected in the deviation and dispersion of the gray level at a certain point of the image from the overall average value. It is defined as:

$$SD = \sqrt{\frac{1}{m \times n} \sum_{x=1}^{m} \sum_{y=1}^{n} (F(x,y) - \bar{F})^2}, \tag{20}$$

where $\bar{F}$ is the pixel mean value of the fused image. The value of $\bar{F}$ can be obtained by the following formula:

$$\bar{F} = \frac{1}{m \times n} \sum_{x=1}^{m} \sum_{y=1}^{n} F(x,y). \tag{21}$$

Average Gradient (AG). The average gradient mainly reflects the small local texture information of the image, which is used to measure the image clarity. The expression is as follows:

$$AG = \frac{1}{m \times n} \sum_{x=1}^{m} \sum_{y=1}^{n} \sqrt{\frac{1}{2}((\partial H)^2 + (\partial V)^2)}, \tag{22}$$

where ∂H is the gradient in the horizontal direction of the image, and ∂V is the gradient in the vertical direction of the image, as shown in the following formula:

$$\partial H = F(x,y) - F(x+1,y), \tag{23}$$

$$\partial V = F(x,y) - F(x,y+1). \tag{24}$$

Mutural Information (MI). Mutual information is used to represent the information transmitted from the source image to the fused image. The definition is as follows:

$$MI = \sum_{k \in \{I,V\}} MI_{kF}. \tag{25}$$

The calculation method of MI_{kF} is as below:

$$MI_{kF} = p_{kF} \log_2 \frac{p_{kF}}{p_k p_F}, \tag{26}$$

where p_k and p_F separately represents the edge histogram of source image X_k and the fused image F, p_{kF} denotes the joint histogram of source image and the fused image.

4.3 Parameters Setting

In the field of image fusion, the suitable parameter settings are important. In our fusion model, the value of the two parameters in the process of image decomposition will have a direct impact on the quality of the fuse image. Based on the experience of Zhang et al. [18], σ_R is set to the default value of 0.1, which is used to control the range weight of the filter and improve the robustness. σ_S determines the clarity of the output image of the filter. The larger the σ_S value is, the better the detail of the output image is, but the worse the texture is. After several experiments, we noticed that base layer and detail layer of the source image will have good visual effect when σ_S =3. The value of ω is taken as 1, the effect is to consider the values of each pixel individually, therefore no details are lost. As shown in Fig. 5, the performance of the filter is tested by several iteration experiments. Finally, the number of iterations is set to 4. At this time, the small structures of the source image are smoothed out, while the edge information is closest to the original image. If increasing the number of iterations continue, the large-scale edge of the image is distorted, which reduces the overall quality of the image.

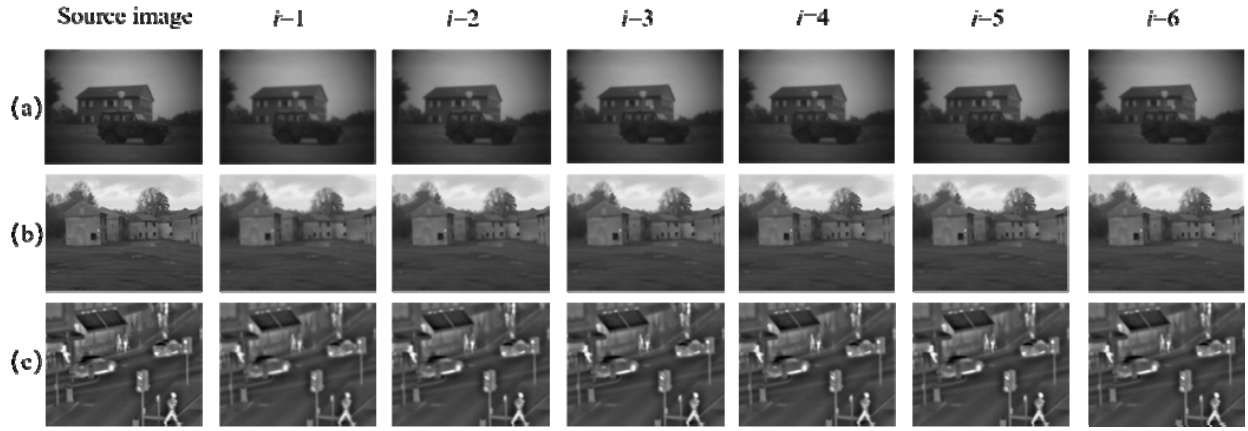


Fig. 5. Output images of three groups of image with different iterations

4.4 Experiment Result and Discussing

Experiments selectes the "field" image pair and "jeep & house"from TNO dataset. The source image and the fusion image of all algorithms are shown in Fig. 6 and Fig. 7. In Fig. 6, the results of RGFF and NSCT methods are good however, infrared information of some regions is missing, and NSCT method produces artifacts. The fused results obtained from VSMWLS and IFEVIP are also satisfactory, but many deep features are still missing. CNN and ResFuse methods can produce better results, but they are still failed in preserving the edges of objects. MDLatLRR performs well in several sensory indicators, but it has the disadvantage of being stiff and unnatural. Apart from these methods, we can see that the fused images produced by our proposed method are free from blur artifacts and it preserves the information of both modalities. In Fig. 7, the contrast of our method is obviously higher, the details of jeep and the windows of the house are obviously reconstructed. We also select the TNO dataset for the quantitative evaluation of the proposed method. In this experiment, ten source image pairs are utilized to obtain fuse images from representative methods and proposed method. Fig. 8 presents ten pairs fused results of VIS and IR images from the TNO dataset. Fig. 9 shows a graphical comparison of five image fusion metrics on 10 fused images. The average scores of the fused results using 10 source image pairs from TNO dataset are shown in Table 1. All indcators are positive. The largest value is shown in bold, the second largest value is represented by a double underline, and the third largest value is shown in single underline, which validates that the proposed approach achieves improved performance over other image fusion techniques.

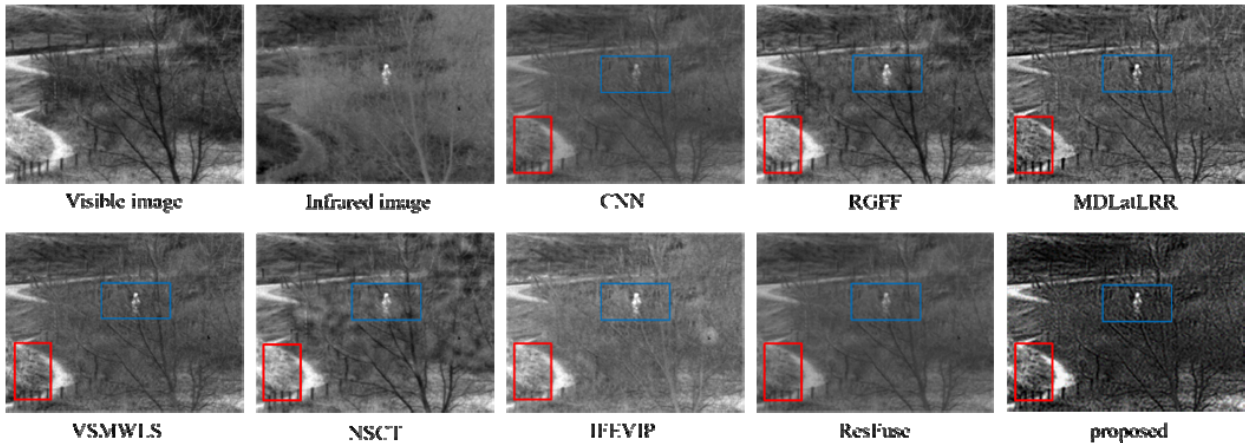


Fig. 6. Visual fused results of 'field' source image pair obtained from different fusion methods

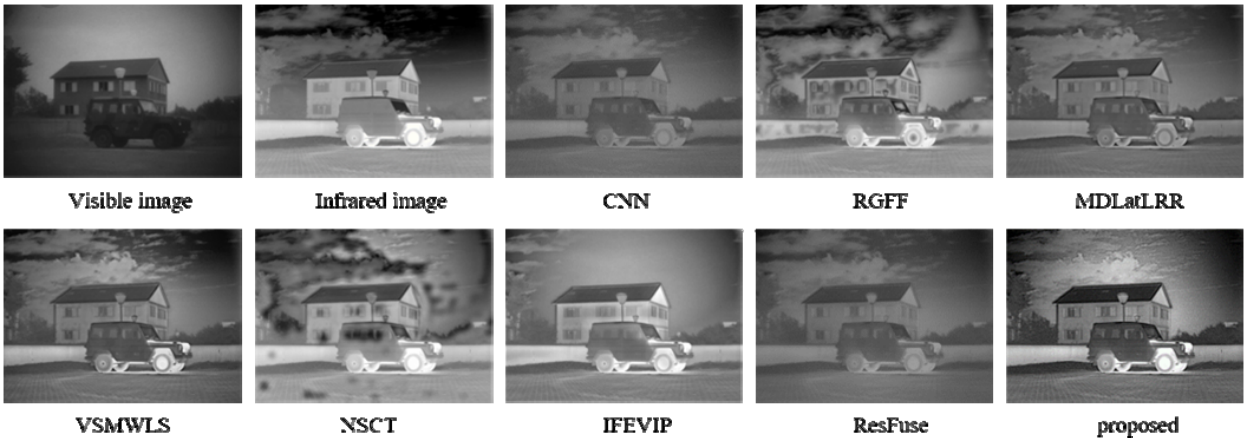


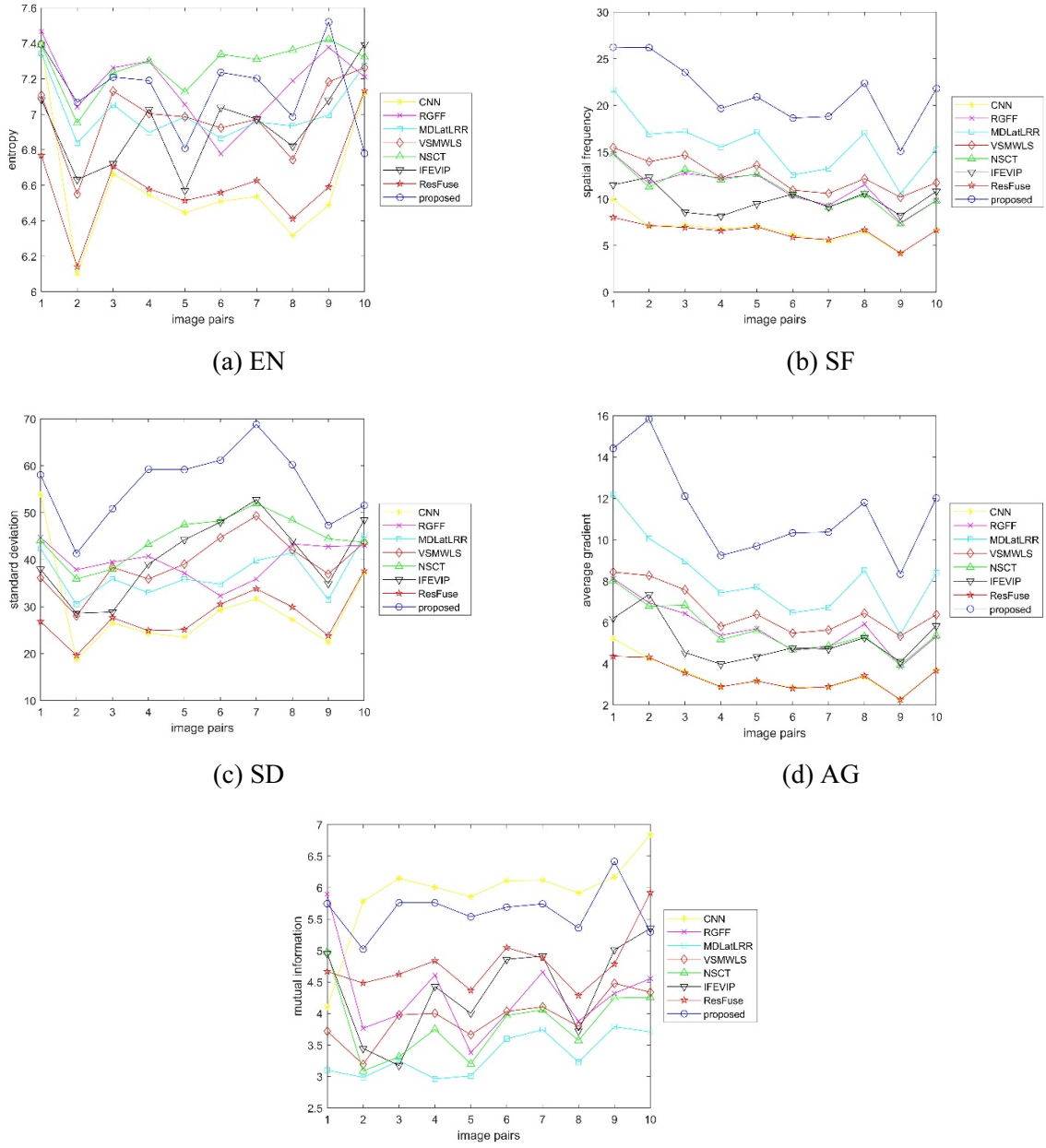
Fig. 7. Visual fused results of 'jeep & house' source image pair obtained from different fusion methods

Fig. 8. Fused results of visible and infrared image pairs from the TNO dataset. (a) Visible images; (b) Infrared images; (c) CNN; (d) RGF; (e) MdLatLRR; (f) VSMWLS; (g) NSCT; (h) IFEVIP; (i) ResFuse; (j) proposed

Table 1. The average quantitative scores of five fusion metrics for 20 fuse dimage obtained from the TNO dataset

Metric	CNN	RGFF	MDLatLRR	VSMWLS	NSCT	IFEVIP	ResFuse	proposed
EN	6.6122	**7.1653**	7.0140	6.9858	**7.2752**	6.9333	6.6024	**7.1392**
SF	6.6871	11.2390	**15.7160**	**12.5555**	11.1049	9.9173	6.4651	**21.3294**
SD	29.4777	39.7205	37.0337	39.4098	**44.5366**	**40.6785**	27.9729	**55.7706**
AG	3.4176	5.7121	**8.1863**	**6.5727**	5.6593	5.1072	3.3313	**11.4161**
MI	**5.9034**	4.3026	3.3366	3.9297	3.8421	4.3852	**4.7885**	**5.6313**

As can be seen from Fig. 9, it is obvious that the traditional method has advantages in the first four evaluation indexes. We notice that the multi-scale transform method, such as NSCT and RGFF, it has good performance in both EN and SD indexes. This shows that the results of multi-scale transformation can retain the main content of the source image. MDLatLRR is also a multi-level decomposition method in essence. It extracts the details and basic parts of the input image at several representation levels by a pre learning projection matrix. MDLatLRR is in second place in both SF and AG evaluation indexes. This is because the gray level of the fusion image changes strongly and the sense of hierarchy is clear. It shows that MDLatLRR method can generate the fused image with obvious contour and edge. The saliency-based method has no short board in each five evaluation indexes, it is still a very competitive integration method. In the fifth index MI, The method of deep learning shows its absolute advantages. CNN and Resfuse are in the top three. Compared with the traditional methods, the deep learning method can obtain the deep details of the source image. The proposed method combines the advantages of traditional methods and deep learning.



(a) EN

(b) SF

(c) SD

(d) AG

(e) MI

Fig. 9. Comparison graphs for quantitative results of different image fusion methods using five image fusion metrics: (a) EN; (b) SF; (c) SD; (d) AG; (e) MI

It can be seen more intuitively from Table 1. The maximum value of each evaluation index is indicated by red, the second largest value is represented by green, and the third largest value is represented by blue. These five evaluation indicators are positive, therefore, the larger the evaluation index value, the higher the fusion quality. The proposed method is ahead of other algorithms in SF, SD, and AG. For EN and MI, the proposed method is also in the top three. It validates that the proposed approach achieves improved performance over other image fusion techniques. Fig. 10 shows the average runtime of different methods on 10 pairs source images. It can be seen that our method does not take much time in the method based on deep learning. This shows that the proposed method's performance is not achieved by increasing the complexity of the algorithm.

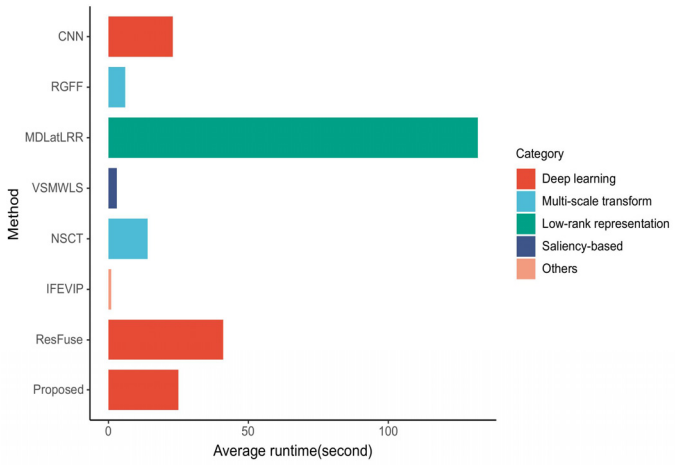


Fig. 10. Average runtime of different methods on 10 pairs source images (unit: second)

5 Conclusions

In this research, a new framework combined rolling guidance filter and VGGnet is proposed. Rolling guidance filter is used for high quality source image decomposition, and VGGnet is used to extract deep details. Experimental results show that the proposed methodmore energy information and richer texture. The feasibility of the algorithm is verified by comparing the source image and the fused image based on the five evaluation indexes. Compared with other representative methods, the proposed algorithm has better fusion performance. In the future, dense blocks can be used to improve the model and enhance the effectiveness of feature extraction.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China under Grant 61801319, in part by Sichuan Science and Technology Program under Grant 2021YFG0099, 2020JDJQ0061, 2020JDJQ0075, 2019YJ0476 and 2020YFSY0027, in part by the Wuliangye project under Grant CXY2020ZR006, in part by the Sichuan University of Science and Engineering Talent Introduction Project under Grant 2020RC33, in part by Innovation Fund of Chinese Universities under Grant 2020HYA04001.

References

[1] J. Ma, C. Chen, C. Li, J. Huang, Infrared and visible image fusion via gradient transfer and total variation minimization, Information Fusion 31(2016) 100-109.

[2] J. Ma, Y. Ma, C. Li, Infrared and visible image fusion methods and applications: A survey, Information Fusion 45(2019) 153-178.

[3] Z. Zhou, B. Wang, S. Li, M. Dong, Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters, Information Fusion 30(2016) 15-26.

[4] J. Chen, X. Li, L. Luo, X. Mei, J. Ma, Infrared and visible image fusion based on target-enhanced multiscale transform decomposition, Information Sciences 508(2020) 64-78.

[5] Q. Zhang, Y. Fu, H. Li, J. Zou, Dictionary learning method for joint sparse representation-based image fusion, Optical Engineering 52(5)(2013) 057006.

[6] Z. Zhu, G. Qi, Y. Chai, H. Yin, J. Sun, A novel visible-infrared image fusion framework for smart city, International Journal of Simulation and Process Modelling 12(2)(2018) 144-155.

[7] N. Mitianoudis, T. Stathaki, Pixel-based and region-based image fusion schemes using ica bases, Information fusion 8(2)(2007) 131-142.

[8] D.-P. Bavirisetti, R. Dhuli, Two-scale image fusion of visible and infrared images using saliency detection, Infrared Physics & Technology 76(2016) 52-64.

[9] X. Huang, G. Qi, H. Wei, Y. Chai, J. Sim, A novel infrared and visible image information fusion method based on phase congruency and image entropy, Entropy 21(12)(2019) 1135.

[10] Y. Liu, X. Chen, H. Peng, Z. Wang, Multi-focus image fusion with a deep convolutional neural network, Information Fusion 36(2017) 191-207.

[11] H. Li, X.-J. Wu, J. Kittler, Infrared and visible image fusion using a deep learning framework, in: Proc. 2018 24th International Conference on Pattern Recognition (ICPR), 2018.

[12] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, Fusiongan: A generative adversarial network for infrared and visible image fusion, Information Fusion 48(2019) 11-26.

[13] K.-R. Prabhakar, V.-S. Srikar, R.-V. Babu, Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs, in: Proc. 2017 IEEE International Conference on Computer Vision (ICCV), 2017.

[14] H. Li, X.-J. Wu, Densefuse: A fusion approach to infrared and visible images, IEEE Transactions on Image Processing 28(5)(2018) 2614-2623.

[15] H. Li, X.-J. Wu, T.-S. Durrani, Infrared and visible image fusion with resnet and zero-phase component analysis, Infrared Physics & Technology 102(2019) 103039.

[16] Z. Gao, Y. Zhang, Y. Li, Extracting features from infrared images using convolutional neural networks and transfer learning, Infrared Physics& Technology 105(2020) 103237.

[17] L. Yan, J. Cao, S. Rizvi, K. Zhang, Q. Hao, X. Cheng, Improving the performance of image fusion based on visual saliency weight map combined with cnn, IEEE Access 8(59)(2020) 976-986.

[18] Q. Zhang, X. Shen, L. Xu, J. Jia, Rolling guidance filter, in: Proc. European conference on computer vision, 2014.

[19] L. Jian, X. Yang, Z. Zhou, K. Zhou, K. Liu, Multi-scale image fusion through rolling guidance filter, Future Generation Computer Systems 83(2018) 310-325.

[20] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014.

[21] A. Toet, J.-K. Ijspeert, A.-M. Waxman, M. Aguilar, Fusion of visible and thermal imagery improves situational awareness, Displays 18(2)(1997) 85-95.

[22] H. Li, X.-J. Wu, J. Kittler, Mdlatlrr: A novel decomposition method for infrared and visible image fusion, IEEE Transactions on Image Processing 29(2020) 4733-4746.

[23] J. Ma, Z. Zhou, B. Wang, H. Zong, Infrared and visible image fusion based on visual saliency map and weighted least square optimization, Infrared Physics & Technology 82(2017) 8-17.

[24] H. Li, H. Qiu, Z. Yu, Y. Zhang, Infrared and visible image fusion scheme based on nsct and low-level visual features, Infrared Physics & Technology 76(2016) 174-184.

[25] Y. Zhang, L. Zhang, X. Bai, L. Zhang, Infrared and visual image fusion through infrared feature extraction and visual information preservation, Infrared Physics & Technology 83(2017) 227-237.

[26] X. Jin, Q. Jiang, S. Yao, D. Zhou, R. Nie, J. Hai, K. He, A survey of infrared and visual image fusion methods, Infrared Physics & Technology 85(2017) 478-501.

[27] S. Klonus, M. Ehlers, Performance of evaluation methods in image fusion, in: Proc. 2009 12th International Conference on Information Fusion, 2009.