Blur Regional Features Based Infrared And Visible Image Fusion Using An Improved C3Net Model

Zhao Xu1,a, Gang Liu1,b*, Li Li Tang1,c, Yan Hui Li2,d

1 School of Automation Engineering, Shanghai University of Electric Power, Shanghai 200082, China
2 Shanghai BrillianceTech Power Equipment & Engineering Co., Ltd, Shanghai 200050, China
a-e-mail: xuzhao025@gmail.com, b-e-mail: superkatherine05@gmail.com, c-e-mail: Liyanhui@brilliance-tech.cn
*Corresponding author’s: b-e-mail: liugang@shiep.edu.cn

Abstract: For ameliorate the drawback that useful information obtained through middle layers is lost in the conventional image fusion methods based on deep learning, an unsupervised deep learning framework based on Cascaded Convolutional Coding Networks (C3Net) is proposed for the fusion of infrared and visual images. A Blur Regional Features (BRF) scheme is also considered during fusion stage, so as to preserve the consistency of regions. Firstly, redundant and complementary features of infrared and visible images are obtained from the coding layer respectively. The output of each convolutional layer is connected to the input of the next layer in a cascading manner. Then, relying on the features of redundant features and complementary features, different fusion strategies are designed respectively based on BRF to obtain fusion feature maps. Finally, the fused image is reconstructed by decoding layer. Furthermore, the objective function of the training model is designed as a multitask loss function including Mean Square Error, Information Entropy and Structural Similarity, to reduce the loss of the original image information. The experimental results of C3Net fusion method is compared with state-of-the-art fusion methods, which is better synthesized performance in objective evaluation and subjective visual quality.

1. Introduction

Infrared and visible image fusion is an significant task in the field of image processing, and the fusion methods of this work have been widely applied in many fields, such as video surveillance, military exploration and overhead power line detection, which is helpful for target detection and recognition. Infrared images can distinguish the target from the background according to the thermal radiation difference, which works well in all weather conditions. Visible images provide texture details with high spatial resolution and clarity in a manner consistent with the human visual system. Therefore, by combining the advantages of thermal radiation information in infrared images and texture details information in visible images, significant features can be extracted from source images, and then these features are fused into an image as complete as possible through appropriate fusion methods[1].

In the past few decades, numerous image fusion algorithms have been developed, among which transform-based algorithms are very popular. However, most of the current image fusion methods use the local fusion operator to carry out the global unified fusion process. The difference of information
characteristics between different regions in the same image is ignored. Therefore, local important information is easily lost. Region-based fusion strategies tend to pay more attention to regional features to achieve good regional feature consistency. Zaveri et al[2] compared fusion methods based on regional features and multi-resolution fusion methods respectively, and verified the importance of local regional feature information for auxiliary decision-making in terms of medical image fusion. Traditional fusion methods consider the problem of regional feature consistency, nonetheless, useful information will still be lost when extracting features, and the source image information cannot be retained effectively.

In recent years, deep convolutional neural network has become an effective image fusion tool, and many fusion methods based on deep convolutional neural network have been proposed successively, which use the depth characteristics of source images to generate the fused image. Liu et al[3] proposed a fusion method based on a convolutional neural network for infrared and visible images, in which image patches that contain different blur versions are used to train a network to generate decision map. The fused image is obtained by decision map and source image. In ICCV 2017, Prabhakar et al[4] proposed a new multi-exposure image fusion framework based on CNN, which extracted the feature information of the source image using the shared weight encoding network, then obtained the fusion feature map using the fusion unit of addition strategy, and finally obtained the fused image through decoder reconstruction. Li et al[5] proposed the DenseFuse fusion framework, which solves the problem of loss of useful information in the middle layer by cascading the encoder and achieves a better fusion effect, but ignores the regional feature consistency of the image. Ma et al[6] applied the generation countermeasure network to the image fusion task, and used an end-to-end fusion framework to directly obtain the fused image. Although the design of complex fusion strategy was omitted, regional details and edge information of target features were also lost.

Addressing the above problems, the new Cascading Convolutional Coding Network (C3Net) is proposed for infrared and visible image fusion based on the DenseFuse model. Firstly, the complementary and redundant features of infrared and visible images are extracted by a coding network. The extracted redundant features and complementary features are fused with different fusion strategies which is softly integrated by Blur Regional Features (BRF) method[7]. The redundancy feature adopts BRF-based feature weighted fusion strategy, while the complementary feature adopts BRF-based feature choose-max fusion strategy. Finally, the fused image is reconstructed by the decoder. The encoding network is composed of four layers of convolution, in which the output of each layer is used as the input of the next layer in a cascading manner to reduce the loss of information in the middle layers. The experimental results of C3Net fusion method is compared with state-of-the-art fusion methods, which is better synthesized performance in objective evaluation and subjective visual quality.

2. Fusion method based on C3Net

2.1 The proposed network architecture
Infrared and visible images are acquired by different sensors from the same scene and registered. They have common scene feature information, which is called redundant features. Owing to the different imaging principles of infrared and visible image sensors, infrared image and visible image contain their own unique feature information, which is called complementary feature.

The cascade encoder is used to extract redundant and complementary features of infrared and visible images, separately. The extracted redundant features and complementary features are fused with different fusion strategies which is softly integrated by Blur Regional Features (BRF) method[7]. The redundancy feature adopts BRF-based feature weighted fusion strategy, while the complementary feature adopts BRF-based feature choose-max fusion strategy. Finally, the fused image is reconstructed by combining the fused features and transmitting them into decoders of sharing weights.
(1) Training network architecture

In the training phase, the training of cascade encoder and decoder is only considered (adding fusion unit in the test phase). Here the cascade encoder and decoder are trained to reconstruct the input image, in order to acquire the complementary and redundant features of the original infrared and visible images (image after strict registration). When extracting complementary features of two inputs, the weights of the two cascade encoder blocks are not shared. Therefore, the features learned respectively are the mapping of complementary relationship between infrared image and visible image. In this manner, the mapping of redundant relation between infrared and visible images can be learned forcibly. In the decoder, the purpose of weight sharing is to combine the complementary features and redundant features of the previous learning, so as to reconstruct the input image. After training the weights of cascade encoder blocks and decoder, adaptive fusion strategy is adopted to fuse the depth characteristic representation obtained by cascade encoder blocks. The advantage of this training method is that appropriate fusion units can be designed for specific fusion tasks. To reduce the loss of source image information, the pooling layer is removed from all convolution blocks in the network. The detailed structure of the training is shown in Figure 1.

![Figure 1. The framework of training process.](image)

(2) Design of multitask loss function

For precise reconstruction of fused images, it is crucial to choose proper loss function that can minimize loss. These common image quality metrics are selected based on two principles: it is helpful to the reconstruction of fused images and improves the quality of fused images. In order to enhance the visual quality of the fused image, and the reconfiguration ability of the fusion network is considered. Nevertheless, Mean Square Error(MSE) and Information Entropy(IE)[8] are computed by reconstructed images and input images, which is conducive to the reconstruction of fused images, and Structural Similarity(SSIM)[9] can improve the visual quality of fused images in the human visual system. So instead, for the sake of making C3Net better learn the complementary and redundant features between infrared and visible images, MSE, SSIM and IE are used as multitask loss functions in the encoding and decoding network. The multitask loss function $L$ is shown in Equation (1).

$$ L = L_{\text{mse}} + \lambda_1 L_{\text{ssim}} + \lambda_2 L_{\text{ie}} $$

where, $L_{\text{mse}}$ is the mean square error loss, $L_{\text{ssim}}$ is the structural similarity loss, and $L_{\text{ie}}$ is the information entropy loss. They are combined with weights $\lambda_1$ and $\lambda_2$ to form the loss function $L$. 

3
2.2 BRF-based fusion strategy

As a programming though in regional fusion, Considering the region consistency fusion, the structure information of the fused image in the target region can be preserved. Especially, the region containing the target information in the redundant information can be effectively fused. Hence, fusion rules suitable for complementary features and redundant features respectively in the feature space of the hidden layer are designed, which is based on the though of BRF method.

(1) The fusion of complementary features

The complementary features of infrared image and corresponding visible image are highly correlated, reflecting the details of the original image. Therefore, the feature selection maximum fusion strategy based on BRF method is adopted to deal with the fusion of complementary features. The fusion complementary feature mapping \( F^C_m(i,j) \) is calculated as:

\[
F^C_m(i,j) = \frac{\sum_{r \in \{1,2,...,R\}} \left(F^\text{IRC}_m(i,j), F^\text{VisC}_m(i,j)\right) \cdot F^\text{IRC}_m(i,j)}{\sum_{r \in \{1,2,...,R\}} F^\text{IRC}_m(i,j), F^\text{VisC}_m(i,j) \cdot F^\text{IRC}_m(i,j)}
\]

where, \( F^\text{IRC}_m(i,j) \) and \( F^\text{VisC}_m(i,j) \) are complementary feature mappings in region \( r \) \( (r \in \{1,2,...,R\}) \) obtained from the source image through the cascade encoder block. \((i, j)\) is the corresponding position of the feature representation and its corresponding fused complementary feature mappings. The fused complementary feature mapping is used as the input of decoder.

(2) The fusion of redundant features

Since the redundant features of infrared and visible images are extracted by sharing weights, they mirror the public information of source images. A feature weighted fusion strategy based on BRF method is proposed. Unlike complementary features, redundant features display complementary relationship when some feature mapping is inactive, but redundancy relationship when some feature mapping is active. In order to better distinguish features, it is a good method to calculate the activity level map by means of L1-norm and average operator of sliding window. \( F^\text{VisR}_m(i,j) \) and \( F^\text{IRR}_m(i,j) \) represent the redundant feature mappings in region \( r \) \( (r \in \{1,2,...,R\}) \) and \( m \in \{1,2,...,M\} \) represents a layer of feature map in the feature map. The average operator of the sliding window is utilized to calculate the final activity level map \( \overline{C'}_r(x,y) \) in region \( r \) by equation (3).

\[
\overline{C'}_r(x,y) = \frac{\sum_{p=1}^{p} \sum_{x=-p}^{p} F(x + \alpha, y + \beta)}{(2p + 1)^2}
\]

where \( p \) is the size of the sliding window. To avoid the loss of some details, set \( p=1 \). The weight coefficient \( w'_i \) is calculated by equation (4).

\[
w'_i = \frac{\overline{C'}_r(x,y)}{\overline{C'}_{\text{VisR}}(x,y) + \overline{C'}_{\text{IRR}}(x,y)}
\]

The fused redundant feature mapping \( F^R_m(i,j) \) is calculated by equation (3), and the fused redundant feature mapping is utilized for the input of the decoder[10].

\[
F^R_m(i,j) = \sum_{r=1}^{R} \left(w'_i \cdot F^\text{VisR}_m(i,j) + w'_i \cdot F^\text{IRR}_m(i,j)\right)
\]

3. Experimental Results and Analysis

3.1 Training analysis
In the training phase, the input image of the network comes from OSU Color-Thermal Database[11]. There are orders of magnitude differences between MSE loss, SSIM loss and IE loss. Therefore, setting different weight coefficients has distinct impacts on the convergence of the network. In the training phase, as shown in Figure 2 (a), when different weights of IE loss and SSIM loss are employed, the total loss presents different convergence. In the first 4000 iterations, the total loss achieved a better value at $\lambda_1=100$, $\lambda_2=10$, when compared to the other three weight settings, followed closely by $\lambda_1=10$, $\lambda_2=100$. Nonetheless, when the iteration is greater than 4000, no matter which loss weight is selected, the C3Net will eventually converge. In the verification phase, the total loss under different weights are capable of evaluating for the network reconstruction capability. As shown in Figure 2 (b), the total loss under different weight settings can reach a better value after 1500 iterations. Thus, it is feasible for C3Net to train under the multi-task loss function.

3.2 Experimental evaluation

For the sake of verifying the effectiveness and superiority of C3Net model, the 24 registered infrared and visible images were selected from the TNO Image Fusion Dataset[12] for performance evaluation. The proposed method is compared with eight typical fusion methods, including WLS[13], CBF[14], LatLRR[15], FPDE[16], CSR[17], CNN[3], FusionGAN[6] and DenseFuse[5]. The codes of these eight comparison algorithms are all from the author's personal home page, and the parameters of each comparison algorithm are set according to the corresponding references. Taking images "Two men in front of house" as examples to evaluate the visual quality of fusion images obtained by eight methods and the methods in this paper using different parameters, as shown in Figure 3.

Figure 2. The graph plot of $Loss$ in training phase and validation phase.

Figure 3. The fusion result of image “Two men in front of house”.
As far as the subjective vision of the human eye is concerned, in the fusion results based on CBF and CSR methods, the fusion images have less significant features and more noises and artifacts, such as branches (blue box) and mailbox and people (red box) in Figure 4. Among the fusion results obtained by WLS, FPDE, CNN, LatLRR and FusionGAN, there are many infrared components in the fusion images, which retain the thermal radiation characteristics, resulting in poor contrast and unclear edge details. Compared with these methods, the fusion results obtained by DenseFuse and this proposed method show relatively high contrast, relatively smooth, clear edge details, rich target information, relatively natural overall visual effect, and better fusion of information in infrared and visible images. The result explains that cascade operation in the middle layer can effectively ameliorate the quality of fused images, and the network trained by the multitask loss function is capable of promoting the quality of the fused images.

In order to compare the fusion results in a more objective and fair way, five quality evaluation indexes were selected to evaluate the fusion results. They are: IE[8]; Mutual Information(MI)[18]; Qabf[19]; the Sum of the Correlations of Differences(SCD)[20]; SSIM[9]. The fusion performance is better with the increasing numerical index of the above five evaluation metrics. According to the 24 fusion images obtained by the above methods, their average values under 5 evaluation indexes are calculated, as shown in Table 1. It is not difficult to see that the results of the proposed method are leading or most leading under the two parameters. Therefore, C3Net model (C3Net-A and C3Net-B) has better fusion performance in infrared and visible image fusion, and the fused images obtained are not only more consistent with human visual perception in qualitative analysis, but also superior in quantitative evaluation.

Table 1. Quantitative comparisons of the five metrics.

|       | IE    | MI(10) | Qabf  | SCD   | SSIM  |
|-------|-------|--------|-------|-------|-------|
| WLS   | 6.5768| 1.3153 | 0.4425| 1.6906| 0.6801|
| CBF   | 6.7649| 1.3523 | 0.4076| 1.4042| 0.5549|
| LatLRR| 6.3056| 1.2611 | 0.4257| 1.6919| 0.6724|
| FPDE  | 6.7331| 1.3466 | 0.4419| 1.6668| 0.6416|
| CSR   | 6.7857| 1.4221 | 0.3257| 1.1964| 0.6221|
| CNN   | 6.8059| 1.4155 | 0.2945| 1.4806| 0.6758|
| FusionGAN | 6.4319 | 1.2864 | 0.2355| 1.5254| 0.6328|
| DenseFuse | 6.7331 | 1.4189 | 0.4647| 1.6667| 0.6913|
| C3Net-A | **6.8309** | **1.4274** | **0.4768**| **1.6973**| **0.7191**|
| C3Net-B | **6.8079** | **1.4237** | **0.4689**| **1.6829**| **0.7034**|

4. Conclusions and Discussion
To further synchronously improve the information content and visual quality of the fused images, the new Cascading Convolutional Coding network (C3Net) for infrared and visible image fusion is proposed. Through the subjective and objective comparison of the different fused images obtained by C3Net and other eight state-of-art methods, it shows that the C3Net proposed in this paper is more effective for the fusion of infrared and visible images. The final C3Net fused image shows that, the thermal target is highlighted, the background details are clearer and more visible and the regional consistency of fused images is more obvious, which is more conducive to the detection and identification of infrared thermal target.

5. Conflict of interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
Acknowledgments
The authors gratefully acknowledge the financial supports by the National Science Foundation of China under Grant number 61203224, as well as Shanghai Pujiang Talent Program under Grant number 16PJJD028.

References
[1] Ma, J, Y; Ma, Y. (2019) Infrared and visible image fusion methods and applications: A survey[J]. Information Fusion, 45(1): 153-178.
[2] Zaveri T; Zaveri M. (2009) Region Based Image fusion for detection of Ewing Sarcoma[C]. In: International Conference on Advances in Pattern Recognition. Kolkata. pp. 240-243.
[3] Liu, Y; Chen, X; Cheng, J. (2018) Infrared and visible image fusion with convolutional neural networks[J]. International Journal of Wavelets, Multiresolution and Information Processing, 16(3).
[4] K. R. Prabhakar; V. S. Srikar; R. V. Babu. (2017) DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 4724-4732.
[5] Li, H; Wu, X. (2019) DenseFuse: A fusion approach to infrared and visible images. IEEE Trans. Image Process. 25: 2614–2623.
[6] Ma, J; Yu, W; Liang, P; Li, C; Jiang, J. (2019) FusionGAN: A generative adversarial network for infrared and visible image fusion. Information Fusion. 48: 11–26.
[7] Liu, G; Jin, Z; Sun, S. Y. (2006) Multiresolution image fusion scheme based on fuzzy region feature[J]. Journal of Zhejiang University-SCIENCE A. 7(2):117-122.
[8] Aardt, Van; Jan. (2008) Assessment of image fusion procedures using entropy, image quality, and multispectral classification. Journal of Applied Remote Sensing.2(1): 1-28.
[9] Wang, Z; Bovik, A, C; Sheikh, H, R; Simoncelli, E, P. (2004) Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process.13(4): 600–612.
[10] Liu, Y; Chen, X; Peng, H; Wang, Z. (2017) Multi-focus image fusion with a deep convolutional neural network. Information Fusion. 36: 191–207.
[11] Davis, J, W; Sharma, V. (2007) Background-subtraction using contour-based fusion of thermal and visible imagery[J]. Computer Vision & Image Understanding. 106(2-3): 162-182.
[12] TNO Image Fusion Dataset. (2018) Available online: https://figshare.com/articles/TN_Image_Fusion_Dataset/1008029.
[13] Ma, J; Zhou, Z; Wang, B; Zong, H. (2017) Infrared and visible image fusion based on visual saliency map and weighted least square optimization[J]. Infrared Physics & Technology. 82: 8-17.
[14] Kumar, S; B, K. (2015) Image fusion based on pixel significance using cross bilateral filter. Signal Image & Video Processing. 9(5): 1193–1204.
[15] Liu, G; Yan, S. (2011) Latent low-rank representation for subspace segmentation and feature extraction. International Conference on Computer Vision. pp. 1615–1622.
[16] Bavirisetti, Durga Prasad; Xiao, G; Liu, G.(2017) Multi-sensor image fusion based on fourth order partial differential equations. International Conference on Image Fusion. pp.1–9.
[17] Liu, Y; Chen, X; Ward, R; Wang, Z. (2016) Image fusion with convolutional sparse representation. IEEE Signal Process. Lett. 23: 1882–1886.
[18] Zhang, X, L; Li, X, F; Li, J. Validation and Correlation Analysis of Metrics for Evaluating Performance of Image Fusion. Acta Automatica Sinica, 2014, 40(02), 306–315.
[19] Xydeas, C S; V, P V. (2000) Objective image fusion performance measure[J]. Military Technical Courier. 56(4), 181-193.
[20] Aslantas, V; Bendes, E. (2015) A new image quality metric for image fusion: the sum of the correlations of differences. AEU: Archiv fur Elektronik und Ubertragungstechnik: Electronic and Communication. 69(12): 1890–1896.