Gated Fusion of Infrared and Visible Light Images Based on CNN

Dong Liu¹, Huihua Yang¹, Lemeng Wang¹, Yuying Shao² and Peng Peng²

¹School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
²Training Center, State Grid Shanghai Municipal Electric Power Company, Shanghai, China

Abstract. As a new research direction, fusion image technology has attracted more and more attention in many fields. Among them, infrared image and visible image, the two kinds of multimodal data have strong complementarity, the fusion image of the two modes contains not only the radiation information of infrared image, but also the texture detail information of visible image. In this paper, a convolutional neural network-based encoding-fusing-decoding network model structure is used. In the encoding stage, Dense Block, which has the advantage of feature extraction, was adapted to extract the image features. In the fusion stage, four fusion methods were compared and analyzed, and the gated fusion was selected as the main method of fusion layer. In the decoding stage, RDB (Residual Dense Blocks) was used to restore the fused features to the fused image. The fusion image based on this method is sensitive to temperature characteristics and has a better performance in image quality. The fused image has a high contrast, a relatively smooth fusion effect, and the overall visual effect is more natural.

Keywords. Infrared image; visible light images; image fusion.

1. Introduction

In the field of image processing, the image obtained by a single sensor can only have a single advantage, and there are some disadvantages. In order to meet the increasingly powerful market demand, image fusion technology should be born. Image fusion is an enhancement technology, simultaneous interpreting images acquired from different sensors to generate more robust or richer information for subsequent processing or decision making.

Many traditional methods have been proposed [1-8] and been proved successful in image fusion. Laplacian pyramid [1] and wavelet transform [2] is classical methods and have been widely adopted for image fusion. To fuse different kinds of images, CNN-based methods have been proposed in recent works [3-5, 8]. For example, Prabhakar et al. [5] proposed an exposure fusion method for extreme exposure image pairs. CNN algorithm was used to fuse the brightness channels of exposure image pairs, and the weighted fusion strategy was used to fuse the chroma channels of input images.

In this paper, we propose a convolutional neural network (convolutional neural network) based code-fush-decoding network model structure. In the encoding stage, DenseBlock, which has the advantage of feature extraction, is used to extract the image features. In the fusion stage, the four fusion methods were compared and analyzed, and the gated fusion was selected as the main method of fusion layer. In the decoding stage, RDB (Residual Dense Blocks) was used to restore the fused features to the fused image.
2. Related Work

There are some methods that related to our work fusing infrared and visible light images. In the last few years, CNN-based methods have already become the most important methods for image fusion. Current multi-sensor image fusion methods have high computational complexity and are difficult to be implemented in hardware. Mishra et al. [9] proposed a fusion method based on two-scale decomposition and improved Frie-Chen operator, which can be implemented in hardware. The architecture proposed in this paper has one throughput unit per clock cycle and is capable of processing 30 HD images per second.

Image fusion algorithms based on traditional deep learning may lose important information acquired by the middle layer. To address the above shortcomings, Zhao et al. [10] proposed an unsupervised deep learning framework based on Cascading Convolutional Coding Network (C3NET), which is applied to infrared image fusion and visible image fusion. In order to maintain regional stability, fuzzy region feature (BRF) programming is also considered in the fusion stage. Ren et al. [11] proposed an image fusion method based on decomposition and partition strategies. This method improves the guide filter, which can achieve good results in image decomposition and limit the artifacts around the image boundary. Since there are more similar images in the vertical direction, the team divided the infrared and visible images into multiple subimages in this direction, with each word image discrete into a base level and a detail level. In order to achieve adaptive output of the fused image, the team also proposed a new fusion strategy called gradient brightness criterion. Based on multi-scale transformation and norm optimization, Li et al. [12] proposed an infrared and visible image fusion method. This method combines L2 norm and L1 norm to design sparse constrained loss function, and optimizes the loss function by splitting Bregman to obtain the final fusion base.

3. Technical Approach

3.1. Overall Architecture

In this part, we will introduce our image fusion network, which is a fusion of visible light images and infrared images. The overall architecture of the proposed network is illustrated in figure 1.

Our proposed visible light image and infrared image fusion network can be divided into three stages: encoding stage, fusion stage, decoding stage. In the encoding stage, DenseBlock, which has the advantage of feature extraction, is used to extract the image features. In the fusion stage, four fusion methods were compared and analyzed, and the gated fusion was selected as the main method of fusion layer. In the decoding stage, the concept of Residual density network was introduced. The Residual density block (RDB) used in it combined the characteristics of Residual block and Dense block, and the two network structures were combined to further improve the network expression ability. The fusion image based on this method is sensitive to temperature characteristics and has a better performance in image quality.
We use feature level image fusion to extract effective features from various image sources, convert the original image data into an expression containing more information, and use it for subsequent applications after comprehensive processing. The significance of infrared and visible image feature fusion is that the natural complementarity between different modes is conducive to the comprehensive description of the target, so as to improve the detection performance. The difficulty of feature level image fusion lies in the incompatibility between different features; How to measure the complementarity between features; How to design reasonable feature level fusion rules.

3.2. Encoding Stage
In this paper, three Dense blocks were designed. C11 and C12 are used to enlarge the input image channel in the encoder. In the Dense Block, the feature maps of each layer are of the same size and are connected through the channel dimension. Among them, the growth rate is a super parameter, set to 16. That is, after convolution of each layer in each Dense Block, the output result is 16 feature maps, that is, the output feature graph channel is 16, which can achieve better performance. With the increase of the number of layers, the channel of each subsequent layer is the sum value of all previous layers.

3.3. Feature-Level Fusion Stage
Considering the complementarity of infrared image and visible image, the infrared image is less affected by illumination and the rich texture details in visible image, several fusion structures of visible and infrared image fusion are designed in this paper.

According to whether the image feature fusion is in a single mode, we divide the image feature fusion into inter-modal fusion and intra-modal fusion.

• Input fusion: As shown in figure 2a, before the infrared image and visible image are input into the network, channel superposition is carried out for them. Due to the increase of input channels, only the first convolution layer needs to be modified accordingly.

• Early fusion: as shown in figure 2b, the feature map and channel of the first dense block in the two backbone networks are first superimposed, and then dimensionality is reduced. Because the resolution of the shallow features of the image is slightly higher than that of the deep features, the fusion method of the shallow features of the image has less convolution, lower semantics and more noise.

• Late fusion: As shown in figure 2c, late fusion is a deep fusion, which overlaps the last convolutional layer channel in the branch of infrared image and visible image. The purpose of the late fusion is to fuse the deep features with stronger semantic information, but it will lead to low resolution and poor detail perception ability.

• Gated Fusion: As shown in figure 2d, gated fusion is an adaptive feature fusion strategy that extracts different features of features and performs weighted summation operations on features. Aiming at the problem that visible light images and infrared images have different contributions to target detection tasks, a fusion strategy based on gate fusion is used for image fusion.

3.4. Decoding Stage
In the decoding stage, RDB (Residual Dense Blocks) was used to restore the fused features to the fused image. This stage consists of three parts: RDBs, DFF, Upsample layer.

RDBs: Residual block (figure 3a) combined with the Residual block (Dense block (figure 3b) to form a Residual block (figure 3c). The convolution operation after concat in RDB is helpful to the training of RDB modules with a larger growth rate.

DFF: This module is composed of integral residual learning and integral feature fusion. The integral feature fusion is the utilization of FI features, and the addition of F-1 and FGF is in the integral residual learning part.

Upsample: This module is the final upsampling and convolution operation of the network, which is used to enlarge the input images.
4. Experiments and Results

4.1. Experimental Environment

The experimental environment of this paper is as follows: the computer processor is Inter(R)Core(TM)i7-8700K CPU@3.70GHZ*12, the image processor is GeForce GTX1080, the operating system is 64-bit Ubuntu 16.04 LTS, and the memory size is 15.5GB. MS-COCO2014 dataset training network was used to perform Image Fusion on TNO Image Fusion dataset, and compared with other methods.
4.2. Training Detail
We designed the model using PyTorch as the framework for training. The backbone network is the Densenet pre-trained on ImageNet. Random Gaussian distribution is used to initialize other parameters of the network, and Adam is selected as the optimization algorithm. Trained on the MS-COCO dataset, the batch size was 16. The initial learning rate was set as, and the learning rate was reduced to after 10K training cycles. The training stopped when the overall loss function changed little.

4.3. Valuation and Comparison
Subjective evaluation means that the tester directly evaluates the image quality, and the result is more simple and intuitive. However, in the process of subjective evaluation, the evaluation results are often affected by human factors, such as the professional degree of testers, psychological aesthetics and testing motivation.

Objective Evaluation: The objective evaluation relies on the real fixed data for testing, which has low cost and is easy to realize, and can be combined with the application system. Objective evaluation method because of the actual existence of numerical standards, whether the image is good or bad can only produce one result, not affected by human factors.

- EN (entropy) is generally used to represent the richness of image information.
- QABF (Fusion Quality) is an objective non-reference quality evaluation index for the relatively novel fusion image evaluation.
- SSIM (Structural Similarity) is a widely used measurement standard at present.

The source image is "UN CAMP" infrared and visible light sequence map taken by TNO Human Factors Research Institute in the Netherlands. The Fusion results are shown in figure 4. In terms of the human eye observation of subjective visual fusion image generated based on convolutional neural network method has a more outstanding performance in image quality, compared with other methods, it can be seen that this algorithm fusion of next generation figure appear sensitive to temperature characteristics, image contrast is higher, a relatively smooth fusion effect, the overall effect is relatively natural visual observation, Visible for infrared and visible image information fusion more thorough. As shown in table 1, the fusion effect has been improved to some extent. Our algorithm model performed better than other fusion-level methods in EN and QABF and better than pixel-level methods in SSIM. In the end, the decoding module uses the super-resolution residual-density network, which makes the fused image have sharper and clearer edges and enrich the target information.

![Figure 4. Fusion results of gray image under different fusion strategies.](image-url)
Table 1. Quality standard of image fusion.

| Method            | EN  | Qabf | SSIM  |
|-------------------|-----|------|-------|
| Pixel-level       |     |      |       |
| NSST-APCNN        | 6.98| 0.65 | 0.69  |
| MSST-FL           | 6.99| 0.53 | 0.70  |
| SCM-CST           |     | 0.55 | 0.69  |
| Feature-level     |     |      |       |
| CNN               | 6.81| 0.29 | 0.71  |
| DenseFuse         | 6.68| 0.44 | 0.73  |
| Ours              | 6.84| 0.48 | 0.72  |

5. Conclusion
The results show that the method of feature extraction and feature fusion used in this paper has achieved excellent results in image evaluation, especially the application of residual-density network has significantly improved the image effect. The innovation of this paper is to select densenet network with dense connection in the coding and fusion stage for multimodal data with strong complementarity, which can not only reduce parameters and improve efficiency, but also obtain rich image features and excellent fusion effect. In the decoding stage, the concept of RDN is introduced to make full use of all the hierarchical features of the original low resolution image.

Acknowledgment
This work was financially supported by State Grid Corporation of China No. 5400-201917455A-0-0-00. Special thanks go to Mr. Yang for his guidance on this model.

References
[1] Vanmali A V and Gadre V M 2017 Visible and NIR image fusion using weight-map-guided Laplacian-Gaussian pyramid for improving scene visibility Sādhanā: Academy Proceedings in Engineering Science 42 1063-1082.
[2] Han X, Zhang L and Du L 2015 Fusion of infrared and visible images based on discrete wavelet transform Sel. Pap. Photoelectron Technol. Committee Conf doi:10.1117/12.2216054.
[3] Li H and Wu X 2019 DenseFuse: A fusion approach to infrared and visible images IEEE Trans. Image Process. doi:10.1109/TIP.2018.2887342.
[4] Liu Y and Wang Z 2014 Simultaneous image fusion and denoising with adaptive sparse representation Image Process. Lett. 9 347-357 doi:10.1049/iet-ipr.2014.0311.
[5] Prabhakar K R, Srikar V S and Babu R V 2017 DeepFuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs, IEEE International Conference on Computer Vision (ICCV) pp 4724-4732 doi:10.1109/ICCV.2017.505
[6] Ma J, Ma Y and Li C 2019 Infrared and visible image fusion methods and applications: A survey Information Fusion 45 153-178.
[7] Zhou Z, Wang B and Li S 2016 Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with gaussian and bilateral filters Information Fusion 30 15-26.
[8] Ma J, Yu W and Liang P 2019 FusionGAN: A generative adversarial network for infrared and visible image fusion Information Fusion 48 11-26.
[9] Mishra A, Mahapatra S and Banerjee S 2017 Modified Frei-Chen operator-based infrared and visible sensor image fusion for real-time Applications IEEE Sensors Journal 4639-4646.
[10] Ren L, Pan Z and Cao J 2021 Infrared and visible image fusion based on weighted variance guided filter and image contrast enhancement Infrared Physics and Technology 114.
[11] Xu Z, Liu G and Tang L 2021 Blur regional features based infrared and visible image fusion using an improved C3Net model Journal of Physics: Conference Series 1820 (1).
[12] Li G, Lin Y and Qu X 2021 An infrared and visible image fusion method based on multi-scale transformation and norm optimization Information Fusion (prepublish).