Recognition of Fusion Processing Based on Infrared and Visible Images under the Framework of Semi-supervised Learning

Zhuohang Cao 1, Shuhao Bian 2, Zhe Chen 3, Peili Ma 2, Shuaishuai Zhai 1 and Lijun Xu 3

1 School of Mechanical Engineering, Nanjing Institute of Technology, Nanjing, Jiangsu, 210000, China
2 School of Automation, Nanjing Institute of Technology, Nanjing, Jiangsu, 210000, China
3 Institute of Art and Design, Nanjing Institute of Technology, Nanjing, Jiangsu, 210000, China

Abstract. In traditional convolutional networks, due to the lack of adequate information protection of traditional image fusion technology and the incomplete removal of redundant noise, the useful information of the fusion image is missing and the recognition success rate is low. In this paper, through the research of deep learning-based image fusion methods and traditional target recognition and SVM neural network, an image fusion processing recognition method based on infrared and visible light is designed. The coding network of this image fusion method consists of convolutional layers, fusion layers and dense blocks. The output of each layer needs to be connected to the next few layers by using a densely connected neural network, so as to obtain more useful features from the source image and fuse the data of the two images better. It is verified by simulation that the fused image has sound visual effects, and its edges and details have been completely preserved. Thus, the target object has strong recognizability compared with the surrounding environment. Research shows that this method will help to more accurately interpret target information in complex environments and achieve more effective results in target recognition.

1. Introduction

1.1. Image fusion processing concepts

Image fusion processing refers to the predetermined processing of different imaging results obtained by multiple sensors on the same target, to aggregate into a new image with more precise information and more excellent content, so as to meet the needs of practical use.

As a classic problem in the field of image fusion processing, infrared and visible image fusion requires us to extract salient features from the source image, and then integrate these features into another target image through an appropriate fusion method [1], and finally output the fusion. The result is used to supplement the missing information for the target image.
1.2. Research progress of image fusion processing methods

In the past few decades, many signal processing methods have been applied to image processing fusion tasks to extract the salient features of source images, such as methods based on multiscale decomposition [2,3]. First, the salient features of the image are extracted by using the image decomposition method, and then the final fusion image is obtained by using a suitable fusion strategy. This method of fusion of infrared and visible light is widely used in video surveillance and military fields and has achieved good results. Besides, image-based learning methods based on representations have also attracted much attention. Among them, the fusion method based on sparse representation (SR) and gradient histogram (HOG) [4] and a low-rank representation (LRR) fusion method [5] has been widely discussed in the industry.

The method of using Convolutional Neural Network (CNN) to obtain image features and reconstruct the fused image is widely used in various literature [6, 7]. However, compared to the methods used in this article, the above CNN-based fusion methods only have the results of the last layer as the features of the image, which will cause the middle layer to lose much useful information obtained from the source image, leading to a poor effect of fusion. For example, in 2016, a deep learning-based fusion algorithm was proposed. The Liu Yu team proposed a fusion method based on sparse convolutional representation (CSR) based on CNN [8]. The effect is shown in Figure 1. Compared with CNN, the CSR extracts the Deep Multi-Layer features of the image and generates a fused image based on these features. Then, a CNN-based joint image fusion and super-resolution method [9] and a convolutional sparse representation method for image fusion [10] also appeared.

![Figure 1. Effect of CSR image fusion method](image)

1.3. Image fusion processing at home and abroad

1.3.1. Face Recognition

In 2018, the US Army Research Laboratory Benjamin S. Riggan proposed a new image synthesis method that uses a convolutional neural network and a series of algorithms to synthesize infrared face images into visible light images. Finally, local and global optimization is performed to produce a synthetic visible light image. The synthesized image is not a perfect visible light image, and it looks more like a "sketch".

1.3.2. Night Vision System

In 2018, NEC and the research team of Professor Tomi Masaaki of Tokyo Institute of Technology, National University of Japan and Masahiro Tanaka jointly developed the "multi-mode image fusion technology". The visible-light image is automatically and efficiently synthesized with the non-visible light image captured by the thermal imaging camera, and the visual recognition of a single image is improved.

In 2019, in order to break through the night imaging effect, Huawei adopted intelligent full-color image fusion technology. This technology uses the intelligent full-color image fusion technology of the AI algorithm to intelligently full-color fusion of infrared transparent black and white images and color images brought by a small amount of visible light, thereby obtaining a clear and bright full-color image.
Through intelligent exposure technology, invisible fill light technology, etc., it has achieved 24h high-definition.

1.3.3. Monitoring System In 2019, Hikvision thermal imaging dual-spectrum heavy-duty gimbal based high-precision pyrotechnic recognition algorithm based on deep learning, the pyrotechnic recognition rate can reach 70%, and the false-negative rate can be reduced to 1 %. Thermal imaging is less affected by the environment. Visible light monitoring supports laser fill light. The fill light distance is 3km, and the minimum illumination is 0.002Lux @ (F1.2, color, AGC ON), which can genuinely realize all-weather monitoring. In particular, the observation effect of thermal imaging at night is the same as or better than that of daytime. It is found that the fire alarm time is earlier than other monitoring methods.

2. Specific implementation methods
First of all, this paper uses the above image fusion method to form a fusion of gray-scale visible light image and infrared image, and finally outputs the fused image. Then, using traditional image noise reduction processing and SVM neural network, the target area is detected.

2.1. Visible and Infrared Image Fusion
First, this article is based on the gray-scale visible light image and infrared image to generate a fused image. After the input image is obtained, the input image is matched according to the method described in [11] for preprocessing, and then passed to the network framework. The framework is divided into three parts: encoder, fusion layer, and decoder (as shown in Figure 2 Show). As can be seen from Figure 2, the encoder is divided into two parts. One is C1 convolution filtering, and the other is DenseBlock. For DenseBlock, we can see from the figure that it contains three convolutional layers. At the same time, it is in a densely connected state, and each output is assigned to one or more inputs.

Figure 2. Data fusion generation model

For the encoder part, the number of input channels for each convolutional layer is 16, and it is passed in and continuously accumulated to ensure that the dense structure of the encoding part can effectively retain the depth feature information of the input source image. This paper uses a unit convolution kernel in the first convolution layer. According to common sense, the convolutional layer should use a small-sized filter with a step size of S = 1. However, in practical applications, smaller step sizes can bring better results. Therefore, if a unit convolution kernel is used, the number of input channels can be reduced, and the parameters and computational complexity of the convolution kernel will be reduced accordingly to achieve rapid dimensionality reduction. Also, this paper uses the same convolution mode. By using a 3 * 3 filter and 1 step size in the encoder, the size of the feature map output after convolution can be maintained. The encoder can be used for input images of any size. In the fusion layer part, the L1-norm strategy is proposed. The L1-norm form is used to calculate the minimum absolute deviation
(LAD), and the weighted graph of the average operator is finally calculated. Using this method can effectively realize the weighted calculation of the parameters of the two images and obtain good results. The final decoder is to obtain the output of the fusion layer and implement decoding to obtain the final fusion image.

2.2. Image Noise Reduction and SVM Neural Network Recognition

After using the unprocessed image to obtain the fusion result, there will be more noise in the picture. Therefore, the traditional image processing noise reduction method is used to reduce the fusion result first, and then Segmentation using SVM.

First, judge the noise points and signal points of the image. For the noise points, count the number of signals in the neighbourhood, and then decide to use the median value of a neighbourhood to replace the noise points according to the number of signals, to remove the salt and pepper noise. The specific implementation steps are as follows: First, the image I is detected, and its pixels I (i, j) are judged. If 0 <I (x, y) <255, the pixel is a signal pixel and saved directly; if I (x, y) = 0 or I (x, y) = 255, the pixel is a noise pixel, then the neighbourhood is counted Number of signals in the network, and look at the value in its neighbourhood. If it is a signal, replace the noise with the value in this neighbourhood. If it is noise, replace the noise with the signal pixels in the neighbourhood. Otherwise, if the number of signals in the neighbourhood is 1, you can expand the neighbourhood window and re-detect; If the number of signals in the neighbourhood is greater than 1, the mean value of the signals in the neighbourhood is used instead of noise. After the above process, the image noise reduction processing with salt and pepper noise is completed.

The SVM is then used to segment the filtered image to find the best segmentation plane from the feature space to maximize the interval between positive and negative samples on the training set. For practical operations, using SVM to solve the binary classification problem and obtaining the required pedestrian area is a relatively simple implementation method. Compared with the effect which other methods such as logistic regression and decision trees achieved (contrast effect as shown in figure 3), the advantages of nonlinear classification of SVM using kernel functions are simply explained from the side. At the same time, neither the logistic regression nor the decision tree model can be fitted with an excellent nonlinear fit. This also caused limitations at the time of the problem.

![Figure 3. Comparison of SVM with other traditional classifications](image)

3. Experimental results and analysis

First, perform the image fusion part of the experiment. During the model training phase, we use the MS-COCO [13] dataset for training. Of these source images, approximately 79,000 images were used as input images. In the verification step, we select 1000 images from MS-COCO as input for training the network. And use pixel loss and SSIM to evaluate the effect. As we can be seen from Figure 4, the SSIM loss increases with time. When the number of iterations is increased to 500, the pixel loss and SSIM get better values when set to larger values. However, when the number of iterations is more significant than 40,000 times, no matter which loss weight is selected, the optimal weight will be obtained. Overall, as the early training phases increase, our network will converge faster. In the training phase, with the larger the time consumption, the less time will be spent.
Figure 4. Pixel loss and SSIM training results for different loss training. And “blue”, “red”, “green”, “yellow” indicate the SSIM loss weight $\lambda = 1,10,100,1000$, respectively.

Then, the effect shown in Figure 5 was achieved. The left and middle are original infrared and grayscale images, and the right is the fused image. The effect is more excellent than before which can effectively capture the corresponding information of the two images. The corresponding information is retained, and the output can be formed very well.

Figure 5. Fusion results of gray-scale visible light and infrared thermal imaging

The next step is filtering the salt and pepper noise on the picture, as shown in Figure 6, the image have less noise, and easy to identify and judge. Compared with the traditional filtering process, this operation has a smaller loss of significant features of the image, and can more effectively and quickly obtain images with low noise and excellent imaging.

Figure 6. Effect before and after noise filtering

Finally, the SVM support vector machine is used to distinguish the target task and the background which can achieve rapid region segmentation and extract the required features. A sparse and robust classifier is used to perform non-linear classification based on the kernel method, and make the effect better than the two methods of linear segmentation and decision tree segmentation. The method of kernel function and SVM is used in this paper. The final result is shown in Figure 7. We fed a certain number of pedestrians from long-range infrared as positive samples, and some buildings, roads, and other images as negative samples. Ultimately, more robust results can be achieved.
4. Conclusion
This paper proposes a method for image fusion processing recognition under the framework of semi-supervised learning. First, a tightly connected neural network structure based on CNN and dense blocks are used to fuse different types of images, and some basic image processing is used to determine and select. Finally, the support vector machine (SVM) classifier performs nonlinear classification to achieve the object selection extraction in mixed images. The result shows that this method can effectively and quickly segment the target area, complete the recognition and classification of objects in the real environment, and effectively reduce the redundant procedures of extracting different features from multiple images, further realizing the need to determine in an image All features. At the same time, this method is also suitable for solving multi-layer image fusion problems, such as multi-focus image fusion, multi-exposure image fusion, and medical image fusion. At the same time, future work will focus on performing new tests in more complex and different environments, generating new public data sets, and exploring new classification methods.

Acknowledgments
This work was financially supported by Major projects of philosophy and social science research in colleges and universities of 2019 Jiangsu province, Research on design of human-machine interaction under the application of AI technology(2019SJZDA118). This work was also financially supported by Innovation and entrepreneurship training program for college students of 2019 in Jiangsu Province, Bionic dolphin design with artificial emotion(201911276012z).

References
[1] Li S, Kang X, Fang L, et al. Pixel-level image fusion: A survey of the state of the art[J]. Information Fusion, 2017, 33: 100-112.
[2] Yang S, Wang M, Jiao L, et al. Image fusion based on a new contourlet packet[J]. Information Fusion, 2010, 11(2): 78-84.
[3] Wang L, Li B, Tian L F. EGGDD: An explicit dependency model for multi-modal medical image fusion in shift-invariant shearlet transform domain[J]. Information Fusion, 2014, 19: 29-37.
[4] Zong J, Qiu T. Medical image fusion based on sparse representation of classified image patches[J]. Biomedical Signal Processing and Control, 2017, 34: 195-205.
[5] Li H, Wu X J. Multi-focus Image Fusion Using Dictionary Learning and Low-Rank Representation[C]//International Conference on Image and Graphics. Springer, Cham, 2017: 675-686.
[6] Liu Y, Chen X, Ward R K, et al. Image fusion with convolutional sparse representation[J]. IEEE signal processing letters, 2016, 23(12): 1882-1886.
[7] Liu Y, Chen X, Peng H, et al. Multi-focus image fusion with a deep convolutional neural network[J]. Information Fusion, 2017, 36: 191-207.
[8] Liu Y, Chen X, Ward R K, et al. Image fusion with convolutional sparse representation[J]. IEEE signal processing letters, 2016, 23(12): 1882-1886.

[9] J. Zhong, B. Yang, Y. Li, F. Zhong. Z. Chen Image fusion and super-resolution with convolutional neural network Chinese Conference on Pattern Recognition (2016), pp. 78-88

[10] Y. Liu, X. Chen, R.K. Ward, Z.J. Wang Image fusion with convolutional sparse representation IEEE Signal Process. Lett., 23 (12) (2016), pp. 1882-1886

[11] Z. Yang, T. Dan, and Y. Yang, Multi-temporal Remote Sensing Image Registration Using Deep Convolutional Features, IEEE Access, vol. 6, pp. 38544-38555, 2018.

[12] Prabhakar K R, Srikar V S, Babu R V. DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs[C]//2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017: 4724-4732.

[13] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//European conference on computer vision. Springer, Cham, 2014: 740-755.