Heterogeneous Image Template Matching Based on Region Proposal Network

Sha Yue¹, Shengzhe Wang²*, Yuyong Cui³, Wei Guan⁴, Xinyi Gao⁵

¹²³⁴⁵Southwest Institute of Technical Physics, ChengDu, SiChuan, China

*email: 1098007659@qq.com
*e-mail: yuesha@uuccb.com

Abstract. Heterogeneous template matching is important to disaster relief. The method of traditional heterogeneous template matching in real-time performance is so weak and usually has a lower accuracy than deep learning method. A heterogeneous template matching method based on region proposal network is proposed in this paper. Specifically, deep features between the optical images and SAR images are extracted by using residual network, and MRPN, a suitable region proposal network for matching work, is proposed to complete precise positioning of the template region. Finally, the channel fusion is performed on the matching results to get a better result and confidence. It is shown that the real-time performance and the accuracy performance have been greatly improved, which is reaching 97%.

1. Introduction

Comprehensive analysis of multi-source images is often used in disaster detection and evaluation in which matching among heterogeneous images plays an important role to obtain a more accurate target area. At present, matching between optical images and SAR (Synthetic Aperture Radar) images is a common mode. A typical matching image is formed through the feature matching of the SAR image and the optical image, helping to discover the hidden features of the target area and to achieve precise disaster area positioning, which has significant meaning in practical application.

Optical images are helpful to people's direct observation, but they can be affected easily by the surrounding environment. SAR images are equipped with the advantage of all day and all-weather, not affected by the environment, which can just make up for the lack of optical images[1]. There are differences in the characteristics of the same features in the image between heterogeneous images because of the different imaging mechanisms. Therefore, difficulties are still in matching issue. Artificially designed operators are used to extract image features in the traditional method, but it is difficult to extract those deep features of heterogeneous images and cause the final matching effect is not good enough. In recent years, better feature extraction and expression capabilities are shown in the deep learning networks and more essential feature information in images could be obtained. Some great breakthroughs are made in the application of deep neural networks of image matching and most of the relevant algorithms are for the processing of homologous image blocks, and the accuracy is higher than that of traditional matching methods. In heterogeneous image matching, there is still a big gap in accuracy demand. Therefore, the research of heterogeneous image template matching methods based on deep learning is becoming much more important.

An end-to-end heterogeneous image matching network is designed by comparing and analyzing the characteristics of heterogeneous images and traditional heterogeneous matching methods. In this paper,
a fusion region proposal network structure which is suitable for matching networks is proposed and a residual structure in the feature extraction network is introduced to make the feature extraction of heterogeneous images more abundant. Finally, the effectiveness of the algorithm in this paper is verified by experiments.

2. Related works
Heterogeneous template matching is mainly divided into traditional feature statistical methods and neural network-based image block matching methods developed quickly in recent years[2]. In 1999, the SIFT [3] matching algorithm proposed by Lowe et al. solved the matching problems caused by illumination, rotation, scale, and other issues. Based on this, many improvements to SIFT in the field of image matching were made by scholars, but the matching results on heterogeneous images were not satisfying. OS-SIFT[4] described the feature information of the image by using the histogram of the similar local gradient direction, but the difference between the gray levels of heterogeneous images was still hard to be avoided completely. Additionally, owing to the application of HOG feature extraction algorithm[5][6] on heterogeneous image matching, the SAR images cannot adapt the coherent speckle noise because of its high sensitivity to noise when promoting the gradient selection. In 2019, RIFT[7], a novel feature matching algorithm robust to large nonlinear radiation distortions was carried out by Li et al., which verified the better matching efficiency of RIFT, compared with SIFT and SAR-SIFT on multimodal images.

With the breakthrough of artificial intelligence technology, the application of deep learning methods in image block matching has been widely used, and the Siamese network has become the mainstream in image matching. One type is a network with a metric layer, like the MatchNet proposed by Han et al[8], equipped with CNN for image region feature extraction and similarity measurement. W Shuang et al.[9] decided to stitch the image block A and the image block B for heterogeneous image matching and to train the VGG network to extract features to point out the differences among the heterogeneous image pairs. Besides the above attempts, one spatial transformation network structure to the block matching neural network and the double-tower structure to match heterogeneous images were introduced by Wang. R C[10]. Another method of deep learning is training the network without a metric layer. The L2-Net proposed by Tian et al.[11] used CNN to convert a batch of image blocks into a batch of descriptors in Euclidean space and applied the nearest neighbor in the batch as the correct matching descriptor. The PN-Net achieved by Balntas et al.[12] contributed to make the minimum negative sample distance pairs larger than positive ones. An end-to-end training matching network LF-Net was firstly presented by Yuki et al.[13] and Shen. X L achieved two improvements based on LF-Net, with an end-to-end image matching network RF-Net[14] based on receptive field.

3. Method

3.1. Network architecture
A network architecture based on region proposal networks applied in template matching between optical images and SAR images is introduced. Thanks to its light-insensitivity and the outstanding anti-interferential ability against smoke and dust, SAR image is applied to be the template image and the optical image becomes the image to be matched, as shown in figure. 1.
The network is mainly divided into a feature extraction network, a fusion area selection network, and a channel fusion fine-tuning module. Among them, a deeper residual network is adopted in the feature extraction network, and the SAR features and optical features are input to the MRPN network after the feature extraction branch. The final output area selection results are fine-tuned for channel fusion after the 3-layer MRPN network, based on which the matching position and confidence estimated of the matching result could be realized.

3.2. Feature Extraction Network

The residual network is applied to the feature extraction network module to complete the target region’s feature extraction. The problem of gradient disappearance and explosion in the training process as the increasement of network layers are solved with the introduction of the residual block, as shown in figure 2. It can be seen from figure 1 that every two residual blocks form layer and there are four layers in total. When extracting features of images, the learning effect could be influenced by the depth of network. It is worth noted that for heterogeneous images, the deeper the network is, the more features at different levels could be learned, and finally, the better matching performance would be carried out.
3.3. Fusion MRPN structure

The RPN (region proposal network) was first shown in Faster-RCNN [15], and then it had been widely used in the field of target tracking, like SiamRPN and SiamRPN++ [16] which proposed in recent years. There are two main branches in RPN network: coordinate regression branch and target classification branch. Coordinate regression branch predicts center point’s offset of position, including center point’s coordinates, length, and width, while target classification branch predicts the target area’s confidence.

Different from the application of region proposal network in target tracking, heterogeneous image template matching has fixed template size, so only the position coordinate offset of the return center point is required on the regression branch. At the same time, one predicted box is generated at one pixel of the correlation map. On the classification branch, a binary label is set for each original box on the graph to be matched whether it is a target or a non-target, and a box that meets the merge ratio above the threshold with the real box is considered a positive sample, as shown in figure. 3. The regression branch only regress on positive samples to simplify the process of the network learning. Based on the above analysis, a matching region selection network MRPN is proposed, whose structure diagram is shown in figure. 4. The SAR image features and optical image features are correlated to obtain a correlation map, and each pixel on the correlation map generates two scores and two coordinate offsets.

Assuming that the center point coordinate’s offsets after regression are $\text{offset}_x, \text{offset}_y$, the corresponding anchor’s coordinates are $\text{anchor}_x, \text{anchor}_y$, and the size of the template image is $w \times h$. The final coordinates on the image to be matched can be obtained by the following equations:

$$x = \text{anchor}_x + \text{offset}_x \times w \quad (1)$$

$$y = \text{anchor}_y + \text{offset}_y \times h \quad (2)$$

The difference between pixels of heterogeneous images is obvious, and finding the common features between heterogeneous images is the key factor to achieve accurate matching. The features extracted from different layers of the neural network are usually different. The feature resolution of the shallow network is high, and the primary features such as texture information are relatively rich, which is conducive to the coarse matching and positioning of the template. The deep network features contain advanced semantic features and guarantee the accuracy of target positioning. Therefore, inserting the MRPN network in different layers of the feature extraction network is more helpful to realize the accurate positioning of the template area on learning different features of the heterogeneous image.
3.4. Fine-tuning module
The result of the region selection network from multiple MRPN layers is passed through a 1*1 convolution kernel to obtain the output result of the final network. In fact, the 1*1 convolution kernel can not only change the dimension, but more importantly, it can merge the information of each channel to a certain extent to achieve cross-channel interaction and information sorting as well. As a result, the final output becomes a fusion of multi-channel information with higher robustness.

4. Experiments and analysis

4.1. Dataset
The dataset used in this experiment is a self-collected dataset, including panchromatic optical remote sensing images and spaceborne SAR images. The optical image size is 800*800 and the SAR image size is 512*512. There are six matching scenes, including ports, cities, rivers, airports, islands, and plains.

4.2. Implementation details
As mentioned in part 3.1, the residual network is used to extract the feature of optical and SAR image separately. The SGD training strategy is applied and the cross-entropy loss function in the classification branch is shown as following:

\[ L_{cls} = -\sum_{i=1}^{N} y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log (1 - \hat{y}^{(i)}) \]  

(3)

In which \( y^{(i)} \) is the classification label and \( \hat{y}^{(i)} \) represents the classification prediction value. Smooth L1 loss function (4) is applied in the regression branch as well and the final loss function \( L \), as shown in Eq(5), needs to be optimized as well.

\[ L_{reg} = \begin{cases} 
0.5x^2 & |x| < 1 \\
|x| - \frac{1}{2} & else
\end{cases} \]  

(4)

\[ L = L_{cls} + \varepsilon \cdot L_{reg} \]  

(5)
In which $\varepsilon$ is the weight adjustment factor.

With the aim of strengthening the final learning effect robustness, data enhancement methods are used, including flipping, shifting, adding noise, changing brightness, etc. The final dataset size is about 80,000 sheets. The initial size of the learning rate is 0.02 and decreases for each epoch, running 50 epochs. The experiment runs on the Windows 10 platform, using the Pytorch framework, the processor is Intel(R) Core (TM) i7-8700 CPU @3.20GHz, and the GPU is NVIDIA GeForce RTX-2060.

4.3. Result on Dataset

Our method results are compared with traditional template matching methods NCC and RIFT on the same dataset, as shown in figure 6, in which the red frame represents the location of the real template area and the green frame is the matching result obtained by the algorithm.

![Figure 6 matching result](image)

The experimental results show that traditional methods and our neural network algorithm both have achieved good matching results on heterogeneous matching image pairs with obvious contour characteristics, but in the case of a large number of similar background interference, like similar buildings in the city, the matching result of the traditional methods are not as satisfying as our method from the visual effect. It is obvious that better performance is presented, and our neural network has certain advantages in the feature extraction of heterogeneous images.

The accuracy of the final matching result is measured according to the IOU value compared with the real location, and the matching could be regarded as a successful one if the IOU is larger than the given threshold.

![Figure 7 Accuracy comparison chart](image)

As shown in figure 7, the comparison of accuracy among different methods tells that when the IoU threshold is 0.3, the matching success rates of the three methods are very high, but with the increase of the IOU threshold, the matching accuracy of the traditional methods, RIFT and NCC, show a downward trend and become much lower than our method. It shows that our proposed algorithm can find the position of the template area more accurately.

| Table 1 Comparison of matching results |
|-----------------------------|-------------------|-------------------|
|                             | RIFT              | NCC               | Ours              |
| MCE                        | 58.719            | 49.62             | 6.7276            |
| IOU>0.9                    | 0.6809            | 0.7               | 0.9787            |
| FPS                        | 6.8               | 1.9               | 9.4               |
As shown in Table 1, the average center pixel error MCE of the NCC algorithm and the RIFT algorithm are large, which indicates that there is a big difference between the matching area and the real area in some scenarios, but our method’s MCE is relatively small. The degree of overlap with the real template area reflects that in different scenarios, traditional methods have certain limitations in feature extraction. In contrast, our method generally has a higher accuracy rate, adapts to different matching scenarios, and has better robustness. Meanwhile, in terms of real-time performance, compared with the high cost of time with traditional methods, our method can process more images in one same period and save time.

5. Conclusion
In this paper, we have introduced a heterogeneous template matching algorithm based on region proposal network and the key idea is to use the residual network to extract the features of optical images and SAR images and use the MRPN to realize the accurate locating of template region. Experimental results demonstrate that the accuracy of our proposed end-to-end network structure is improved by 27% compared with traditional heterogeneous template matching methods.

The method cannot fully adapt to the feature differences among heterogeneous images, like the noise of SAR images has not been eliminated, and the accuracy rate could be further improved. In the future, the model will be improved to extract the features of heterogeneous images according to their respective characteristics.

Reference
[1] Zhu, X.Y. (2016) Research on feature-based SAR and optical image scene matching algorithms. University of Chinese Academy of Sciences, Beijing.
[2] Jia, D., Zhu, N.D., Yang, N.H., Wu, S., Li, Y.X., Zhao, M.Y., (2019) Image matching methods. Journal of Image and Graphics, 24: 0677-0699.
[3] Lowe, D. et al. (2004) Distinctive image features from scale-invariant key points. International journal of computer vision, 91:110.
[4] Sourabh, P., Umesh, C.P. (2018) Optical-to-SAR image registration using modified distinctive order based self-similarity operator. In the IEEE international Students Conference on Electrical, Electronics and Computer Science. Bhopal. pp. 1-5.
[5] Hosotani, D., Yoda, I., Sakaue, K. (2009) Wheelchair recognition by using stereo vision and histogram of oriented gradients (HOG) in real environment. In Workshop on Applications of Computer Vision (WACV). Snowbird, UT. pp. 1-6.
[6] Reichman, D., Collins, L.M., Malof, J.M. (2017) Learning improved pooling regions for the Histogram of Oriented Gradient (HOG) feature for buried threat detection in ground penetrating radar data. In Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXII. Anaheim.
[7] Li, J., Hu, Q., Ai, M. (2018) Rift: multi-modal image matching based on radiation-invariant feature transform. 1-14.
[8] Han, X., Leung, T., Jia, Y., et al. (2015) Matchnet: Unifying feature and metric learning for patch-based matching. In the IEEE Conference on Computer Vision and Pattern Recognition. Boston. pp. 3279-3286.
[9] Wang S, Jiao LC, Fang S, etc. (2018) Heterogeneous image matching method based on deep learning. China, Xi'an.
[10] Wang R C. (2019). Heterogeneous image feature learning and block matching based on deep neural network. Beijing University of Posts and Telecommunications. Beijing.
[11] Tian, Y., Fan, B., Wu F. (2017) L2-net: Deep learning of discriminative patch descriptor in Euclidean space. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 661-669.
[12] Balntas, V., Johns, E., Tang, L.L., et al. (2018) PN-Net: conjoined triple deep network for learning local image descriptors. https://arxiv.org/pdf/1601.05030.pdf
[13] Ono, Y., Eduard, T., Pascal, F., Kwang, M.Y. (2018) LF-Net: learning local features from images[C]. In Advances in Neural Information Processing Systems, 6234-6244.

[14] Shen, X., Wang, C., Zang, Y., Li, X., Li, J., Wen, C., Cheng, M., He, Z. (2019) RF-Net: An End-to-End Image Matching Network based on Receptive Field. In Computer Vision and Pattern Recognition (CVPR). Long Beach. pp. 8132-8140.

[15] Ren, S., He, K., Girshick, R., Sun, J. (2017) Faster R-CNN: Towards real-time object detection with region proposal networks. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 39: 1137-1149.

[16] Li, B., Wu, W., Wang, Q. (2019) SiamRPN++: evolution of Siamese visual tracking with very deep networks[C]. In: Computer Vision and Pattern Recognition. Long Beach. pp. 4282-4291.