GEOMETRIC AND NON-LINEAR RADIOMETRIC DISTORTION ROBUST MULTIMODAL IMAGE MATCHING VIA EXPLOITING DEEP FEATURE MAPS

M. Chen [1, *], Y. Zhao [1], T. Fang [1], Q. Zhu [1], S. Yan [1], F. Gao [2]

[1] Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, 611756, China - minchen@home.swjtu.edu.cn; ytzhao@my.swjtu.edu.cn; 1359465037@qq.com; zhuq66@263.net; shyan@my.swjtu.edu.cn
[2] Shenzhen Real Estate Bid Center, China – gaofeng0006@126.com

Commission III, WG III/6

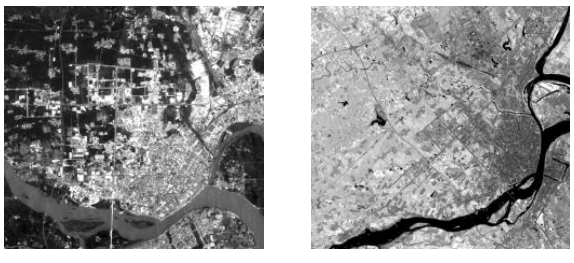
KEY WORDS: Image Matching, Multimodal Images, Geometric Distortion, Non-linear Radiometric Distortion, Deep Feature Maps

ABSTRACT:

Image matching is a fundamental issue of multimodal images fusion. Most of recent researches only focus on the non-linear radiometric distortion on coarsely registered multimodal images. The global geometric distortion between images should be eliminated based on prior information (e.g. direct geo-referencing information and ground sample distance) before using these methods to find correspondences. However, the prior information is not always available or accurate enough. In this case, users have to select some ground control points manually to do image registration and make the methods work. Otherwise, these methods will fail. To overcome this problem, we propose a robust deep learning-based multimodal image matching method that can deal with geometric and non-linear radiometric distortion simultaneously by exploiting deep feature maps. It is observed in our study that some of the deep feature maps have similar grayscale distribution and correspondences can be found from these maps using traditional geometric distortion robust matching methods even significant non-linear radiometric difference exists between the original images. Therefore, we can only focus on the geometric distortion when we deal with deep feature maps, and then only focus on non-linear radiometric distortion in patches similarity measurement. The experimental results demonstrate that the proposed method performs better than the state-of-the-art matching methods on multimodal images with both geometric and non-linear radiometric distortion.

1. INTRODUCTION

Multimodal images reflect different characteristics and information of the observed objects because of the difference of sensor imaging mechanism. Making full use of the complementary advantages of multimodal images can help image interpretation. Image matching, seeking correspondences from overlap image regions, is a fundamental task in the processing and application of multimodal images (Kong et al., 2019; Sedagha and Mohammadi, 2019; Zhang et al., 2019). Because multimodal images are often captured from different platforms, sensors, viewpoints and times, there are significant geometric and non-linear radiometric distortion between multimodal images (see Figure 1), which brings great difficulties to reliable image matching.



(a) (b)

Figure 1. Multimodal images with overlap regions. (a) is a visible band image. (b) is a near-infrared band image. There are significant geometric (scale and rotation) and non-linear radiometric distortion between the two images.

To achieve reliable image matching, many remarkable methods have been proposed. According to the difference of matching strategy, image matching methods can be mainly classified into two categories: area-based method and feature-based method (Gruen, 2012; Chen et al., 2017).

For a pair of multimodal images (one reference image and one target image), the area-based method often sets a window on the reference image and a corresponding search area on the target image, and finds the most similar window in the search area as the matched window. The central points of the two windows are regarded as a pair of matches (Gruen, 2012). The two key points of area-based method are to determine a search area of appropriate size and construct a reliable and robust similarity measurement method. The search area should contain the corresponding window region but not too large. Existing methods usually use direct geo-referencing information or perform a step of coarse registration to roughly eliminate the global geometric distortion between images and then determine a search area of appropriate size. Among similarity measurement methods, the normalized cross correlation (NCC) and mutual information (MI) are robust to radiometric changes to some extent (Chen et al., 2003; Hel-Or et al., 2014). However, they are still difficult to adapt to the non-linear radiometric difference between multimodal images. To improve the matching performance, some methods (e.g. HOPC and CFOG (Ye et al., 2016, 2019)) based on phase congruency model (Kovesi, 1999) have been proposed. However, these methods still have the common problem of other area-based methods: difficult to adapt to image geometric distortion. For example, in

* Corresponding author

HOPC and CFOG methods, image rotation and translation should be coarsely eliminated by direct geo-referencing, and image scale change should be removed based on ground sample distance (GSD). Therefore, they will fail when the prior information of physical sensor models, navigation devices and GSD is unknown or the accuracy is not enough. Another way of eliminating geometric distortion is to find some ground control points (GCPs) to estimate the geometric transformation between images. However, it is difficult to extract reliable GCPs automatically under geometric and non-linear radiometric distortion by using traditional matching methods. Thus, users have to select GCPs manually in many cases. It limits the widespread application of this kind of methods.

Feature-based methods are more robust to image geometric distortion than area-based methods by considering geometric distortion in the designing of feature description algorithm. Feature-based matching methods usually include three steps: feature detection, description and matching. First, feature detectors are adopted to extract features from the reference image and the target image. Interest points, lines and regions are the three most commonly used features. In this paper, feature means interest point if there is no special explanation. Then, a robust feature descriptor is constructed to describe the features. Finally, features are matched based on descriptor similarity measurement. According to the difference of feature detection, description and similarity measurement, feature-based methods can be further subdivided into handcraft methods and deep learning-based methods. Among the handcraft methods, the scale invariant feature transform (SIFT) method (Lowe, 2004) is a milestone work and is widely used in many fields including photogrammetry and remote sensing. With the successful effect of SIFT method, many handcraft methods, such as the speeded up robust features (SURF) (Bay et al., 2008), oriented FAST and rotated BRIEF (ORB) (Rublee et al., 2011), and Affine-SIFT (ASIFT) (Morel and Yu, 2009), have been proposed to improve the performance in time efficiency and robustness to image geometric distortion. To overcome the problems caused by non-linear radiometric difference, some methods on the basis of local self-similarity and phase congruency have been proposed (Huang et al., 2011). However, similar to area-based methods, these methods are not robust to image geometric distortion while improving the robustness to non-linear radiometric distortion.

Recently, with the rapid improvement of computer hardware and software, deep learning technology has attracted more and more attention and has been introduced into the field of image matching (Zagoruyko and Komodakis, 2015; Altwaijry et al., 2016; Melekhov et al., 2017; He et al., 2018). A common idea of deep learning-based matching method is as follows: a deep convolutional neural network with two weights shared branches is constructed and trained on the basis of positive and negative samples by minimizing the distance of deep features between positive samples and maximizing the distance of deep features between negative samples. Studies have shown that deep learning-based matching method is robust to non-linear radiometric distortion between images (He et al., 2019, Quan et al., 2019). However, most of existing deep learning-based matching methods only focus on the non-linear radiometric difference between images. The geometric distortion between images should be coarsely corrected before matching similar to area-based methods.

According to aforementioned introduction, existing matching methods are not robust enough to image geometric and non-linear radiometric distortion simultaneously, which leads to the limitation of practical application. To overcome this problem, we propose a geometric and non-linear radiometric distortion robust multimodal image matching method in the framework of convolutional neural network by exploiting deep feature maps. Firstly, a Siamese-type neural network containing convolutional layers and fully connected layers is designed. For presentation purpose, this network is marked as FSNet (fully connected Siamese-type neural network). FSNet could extract deep features and perform feature similarity measurement. A training dataset is collected by considering negative sample distance to train FSNet. Secondly, the convolutional layers of FSNet are extracted to form another network, marked as CSNet (Siamese-type neural network only has convolutional layers). CSNet is used to produce deep feature maps for multimodal images of any size without known geo-referencing information and GSD. Then, a geometric transformation is estimated by exploiting the deep feature maps to eliminate the geometric distortion between the input multimodal images. After that, interest points are detected from the reference image and the geometric distortion eliminated target image, respectively. Image patches corresponding to interest points are generated and input into the FSNet to find matches. Finally, inliers are recognized from the matching result based on RANSAC method and inversely computed into the original images. The main contributions of this paper are as follows.

1) This paper points out that the main reason that deep learning-based methods can match multimodal images with non-linear radiometric difference is that the non-linear radiometric difference between some deep feature map pairs generated from the last convolutional layer has been alleviated or eliminated. Based on this observation, a multimodal image matching method robust to geometric and non-linear radiometric distortion is proposed. The proposed matching framework is very flexible and can be combined with other advanced image matching neural network in addition to the Siamese-type neural network used in this paper.

2) We find that some non-corresponding image patches with small spatial distances to the corresponding image patch have high similarity because they have large overlapping areas, which leads to mismatches around corresponding points. To overcome this problem, a negative sample generation strategy that takes the distance between non-corresponding and corresponding patches into account is proposed in this paper.

The remainder of this paper is organized as follows. Section 2 presents a Siamese-type neural network and analyses the feasibility of exploiting deep feature maps to deal with the geometric and non-linear radiometric differences between multimodal images. Section 3 describes the proposed multimodal image matching method in detail. The experimental results, along with the method of training dataset generation, matching performance analysis and discussion, are presented in Section 4. The final section concludes this paper and points out possible further improvements that can be made.

2. NEURAL NETWORK CONSTRUCTION AND DEEP FEATURE MAP ANALYSIS

2.1 Network architecture

Siamese-type neural network has been proved to be an outstanding architecture in computer vision tasks in recent years. It has been widely utilized in fields of target tracking, similarity discrimination of images and texts. Because Siamese-type

neural networks perform well for multimodal images without significant geometric distortion, a Siamese-type neural network is designed as the base model of the proposed matching method. This network is marked as FSNet in this paper, as shown in Figure 2. FSNet contains convolutional and fully connected layers but no pooling layer because pooling layers may make the network hard to locate the correct match accurately and finally affect the matching performance. The convolutional layers extracted from trained FSNet form a new network, marked as CSNet. To achieve optimal performance, we tested architectures with different network layers and different convolution kernel sizes. It is found that the network with five convolutional layers (convolution kernel sizes are 3×3, 5×5, 5×5, 5×5 and 5×5) and two fully connected layers performed well.

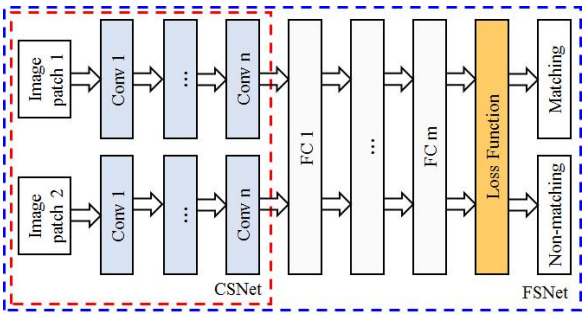


Figure 2. Architecture of the networks used in this paper.

2.2 Loss function

From the perspective of metric learning, image patches matching can be transformed into a binary classification task. A commonly used objective function in classification is cross entropy loss function (Miller et al., 1993). Specifically, the Sigmoid cross entropy loss function (Han and Moraga, 1995) is used in this paper. Given a triplet input (x_i^1, x_i^2, y_i), where (x_i^1, x_i^2) denotes the i_{th} pair of patches in the training dataset and y_i is the corresponding label. The Sigmoid cross entropy loss function can be expressed as Equation (1).

$$\begin{cases} loss = -[y_i \ln p_i + (1-y_i)\ln(1-p_i)] \\ p_i = Sigmoid(\hat{y}_i) = \dfrac{1}{1+e^{-\hat{y}_i}} \end{cases} \quad (1)$$

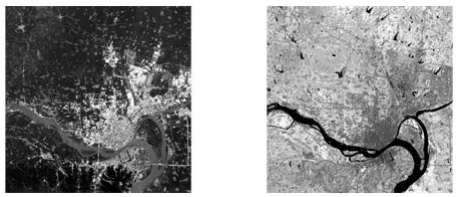
where p_i is the posterior probability of the i_{th} pair of patches, $\hat{y}_i$ is the predicted value of the patch pair (x_i^1, x_i^2). Specifically, the label y_i is expressed as

$$y_i = \begin{cases} 1, & \text{positive patch pairs} \\ 0, & \text{negative patch pairs} \end{cases} \quad (2)$$

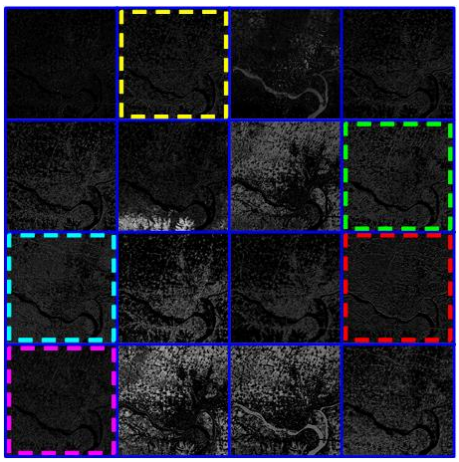
The Sigmoid cross entropy function is used for both clustering the positive samples and separating the negative samples.

2.3 Deep feature maps analysis

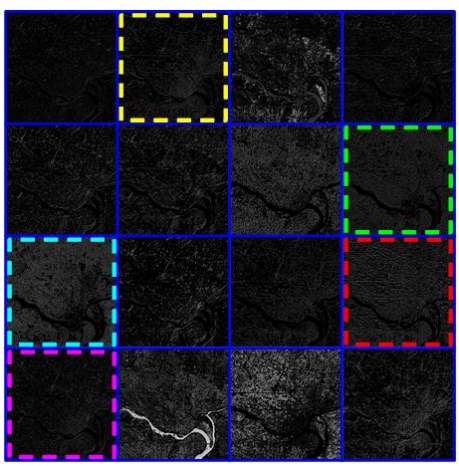
Analyzing the architecture of the network shown in Figure 2, if FSNet can recognize positive sample correctly, we can infer that some of the feature maps output from the two branches of CSNet are similar, that is, there is no significant non-linear radiometric distortion between the corresponding feature maps. This is because the fully connected layers in FSNet mainly play the role of dimensionality reduction and similarity measurement. Figure 3 is an example to demonstrate this conjecture. In Figure 3, the input images are the blue band image and near-infrared band image of a Landsat8 image, respectively. Images shown in the grids are the deep feature maps generated by CSNet. Images in the grids are arranged in the order of the neurons in the last layer of CSNet. It can be seen that although there is significant non-linear radiometric difference between the input images, some deep feature maps that are marked with dotted boxes of the same color have similar appearance. The non-linear radiometric difference has been alleviated or eliminated.



(a) A pair of visible and near-infrared band images



(b) Deep feature maps of the visible band image



(c) Deep feature maps of the near-infrared band image

Figure 3. Deep feature map examples of two images with non-linear radiometric difference. (a) shows the blue band image and near-infrared band image of a Landsat8 image. (b) and (c) show

the deep feature maps generated from the last convolutional layer of CSNet. These deep feature maps are displayed in the order of the neurons in the convolutional layer. Any two images marked with dotted boxes of the same color are a pair of deep feature maps generated by the neurons at the same location in the convolutional layer.

The abovementioned conjecture can be extended to images with both geometric and non-liner radiometric distortion: if a pair of multimodal images of any size with geometric and non-linear radiometric distortion are input into CSNet, the non-linear radiometric difference between some deep feature maps generated from corresponding neurons in the last convolutional layer will be alleviated or eliminated. Under such circumstances, some traditional geometric distortion robust matching methods (e.g. SIFT, SURF or ASIFT) can be adopted to find some correct matches from these pairs of feature maps. And the global geometric distortion between the original images can be eliminated by performing a step of coarse registration. On the basis of this conjecture, a multimodal image matching method that is robust to geometric and non-linear radiometric distortion is proposed in this paper (see Section 3).

3. ROBUST MULTIMODAL IMAGE MATCHING VIA EXPLOITING DEEP FEATURE MAPS

Our goal is to match multimodal images with geometric and non-linear radiometric distortion. In this study, we deal with the geometric distortion and the non-linear radiometric distortion in turn, but the influence of the latter is considered when dealing with the former and vice versa. The flowchart of the proposed robust multimodal image matching method is shown in Figure 4.

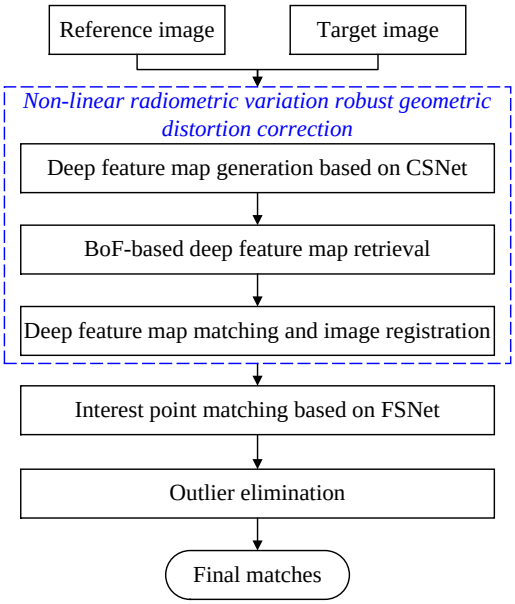


Figure 4. Flowchart of the proposed multimodal image matching method. The steps in the blue dotted box are designed to make the matching method robust to image geometric distortion.

3.1 Non-linear radiometric variation robust geometric distortion correction

3.1.1 Deep feature map generation based on CSNet: According to the analysis in Section 2, the reference image and target image should be input to CSNet to generate deep feature maps. Due to the large size of remote sensing images, especially satellite images, it is very inefficient to input such images directly into CSNet. In order to overcome this problem, the input images are down-sampled at the beginning, and then input into the convolutional layers for processing. In the down-sampling, the down-sampling rate of the reference image and the target image should be kept the same, so as to maintain the fixed scale relationship between the original images. Therefore, the proposed down-sampling method is as Equation (3).

$$\begin{cases} \left[R_r^{'}, C_r^{'}, R_t^{'}, C_t^{'}\right] = \left[R_r, C_r, R_t, C_t\right] / \alpha \\ \alpha = \min\left(R_r, C_r, R_t, C_t\right) / \beta \end{cases} \tag{3}$$

where R_r and C_r are the number of rows and columns of the reference image. R_t and C_t are the number of rows and columns of the target image. $R_r^{'}$ and $C_r^{'}$ are the number of rows and columns of the down-sampled reference image. $R_t^{'}$ and $C_t^{'}$ are the number of rows and columns of the down-sampled target image. α is a scale factor. β is a size factor to determine the size of the down-sampled images, which is empirically set as 600 in this paper.

3.1.2 BoF-based deep feature map retrieval: After down-sampling, the multi-layer convolutional operation is performed to obtain deep feature maps, and the most similar feature map pair can be found. The most straightforward way to find the most similar deep feature map pair and do coarse registration is to perform feature matching on every pair of deep feature maps and the matching result of the deep feature map pair with the most matches is adopted to do coarse registration. However, such a straightforward strategy is inefficient and unreliable. In order to overcome this problem, we use a Bag-of-Feature-based (BoF-based) image retrieval method (Philbin et al., 2007) to measure the similarity of each pair of deep feature maps and select the three pairs with the highest similarity to do feature matching. Because the non-linear radiometric distortion has been significantly relieved between the similar deep feature maps, the SIFT method is adopted in the feature matching.

3.1.3 Image coarse registration: After feature matching, the generated three sets of matches are merged into one group. The RANSAC algorithm is performed on the group to eliminate outliers. Then, a projective transformation is fitted based on the matches. And the original target image is transformed into the coordinate system of the original reference image.

3.2 Interest point matching based on FSNet

Through the process described in subsection 3.1, the geometric distortion between multimodal images has been roughly eliminated. There is only non-linear radiometric difference between corresponding image patches on the reference image and the registered target image. Thus, we use the FSNet to match this kind of patches.

In order to produce image patches for similarity measurement, the Harris detector (Harris and Stephens, 1988) is adopted to detect interest points from the reference image and the registered target image firstly. For each interest point $p_i(x, y)$ on the reference image, a search area of $m \times m$ pixels centred

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-3-2020-233-2020 | © Authors 2020. CC BY 4.0 License.

on (x, y) is determined on the registered target image, and all interest points located in the search area are taken as candidate matches of $p_i(x, y)$. m is set as 30 empirically in this paper. Taking interest point $p_i(x, y)$ as the centre, an image patch of 97×97 pixels is cut from the reference image as reference image patch. An image patch is generated for each candidate match point from the registered target image in the same way. The reference image patch and each candidate image patch are input into FSNet for deep feature extraction and similarity measurement. Among all candidate matches, the candidate with the recognition result "*Matching*" and the highest similarity is regarded as the match point of the reference point $p_i(x, y)$. When all interest points on the reference image have been processed, the RANSAC algorithm is adopted to eliminate outlies and the inliers are computed back to the original image to be the final matches.

4. EXPERIMENTAL RESULTS AND ANALYSIS

In order to demonstrate the effectiveness of the proposed method, we compare it with both handcraft and deep learning-based methods. Among the handcraft methods, the most popular SIFT method (Lowe, 2004) and the multimodal image matching method CFOG (Ye et al., 2019) are selected. Among the deep learning-based methods, the FSNet described in Section 2 is selected as a method of comparison. If the proposed method performs better than FSNet, it will be proved that the proposed matching strategy is effective because FSNet is the base network of the proposed method. All the three compared methods are followed by a step of RANSAC to eliminate outliers as the proposed method.

4.1 Datasets

Three types of multimodal image pairs, including visible-to-infrared, optical-to-SAR, and optical-to-LiDAR are used in our experiments. There is significant non-linear radiometric distortion between images of each pair. To evaluate the robustness of the proposed method to geometric distortion, scale and rotation changes are added to the images manually. Finally six image pairs are formed. The datasets are shown in Figure 5.

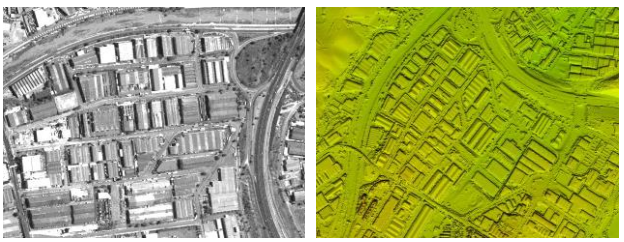
(a) pair 1

(b) pair 2

(c) pair 3

(d) pair 4

(e) pair 5



(f) pair 6

Figure 5. Experimental datasets. (a) and (b) are visible-to-infrared image pairs. (c) and (d) are optical-to-SAR image pairs. (e) and (f) are optical-to-LiDAR image pairs. There is significant non-linear radiometric distortion in all image pairs. Besides, there are scale change between images in pairs 1, 3 and 5. Rotation and scale changes exist between images in pairs 2, 4 and 6.

4.2 Evaluation criteria

In our experiments, two widely used indicators, number of correct matches (NCM) and matching precision (MP), are adopted to evaluate the performance of the proposed matching method. MP is computed as Equation (4).

$$MP = (NCM / NTM) \times 100\% \quad (4)$$

where NTM is the number of total matches. To count the value of NCM, we manually selected some evenly distributed GCPs to fit a projective transformation for each image pair. Then, a

localization error is computed for each pair of match on the basis of the transformation. If the localization error is smaller than a threshold (2 pixels in this paper), the corresponding match is regarded as correct match.

4.3 Training

4.3.1 Training dataset: A training dataset consists of 150000 pairs of matching patches (positive samples) and 150000 pairs of non-matching patches (negative samples) are generated from multimodal images including Google Earth images, ZY3 satellite images, Landsat-8 satellite images, TerraSAR-X satellite images and elevation rendering image of LiDAR point cloud. Multiple land cover types like buildings, rivers, roads and farmlands in urban and rural areas are contained in the training dataset.

In the generation of negative samples, a strategy considering the distance between the centre points of non-matching patches is proposed to overcome the false matching problem caused by neighbour points. As shown in Figure 6, the red patches form a pair of positive sample. Around the corresponding patch (red box) on the target image, eight patches (blue boxes) that are r pixels from the corresponding patch are extracted. One of the eight patches is selected randomly to form a negative pair with the reference patch. We call this kind of negative sample as DNS (distance-based negative sample).

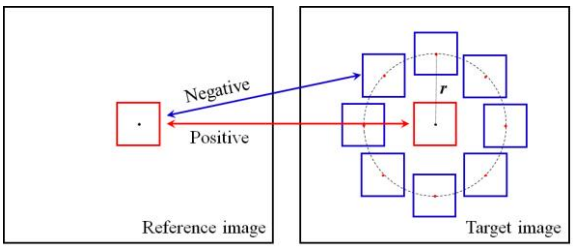
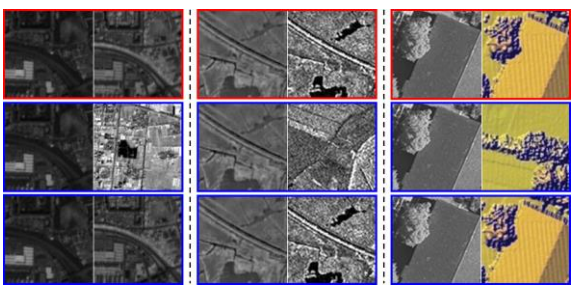


Figure 6. Negative sample generation by considering the distance between the centre points of matching and non-matching patches.

In the training sample generation, if n pairs of positive samples are collected, there will be n pairs of DNS. However, we only select n/2 pairs from all DNS randomly. We produce another n/2 pairs of negative samples randomly from the whole images without considering sample distance. This kind of negative samples is marked as RNS (random negative sample). Therefore, our training dataset contains n pairs of positive samples, n/2 pairs of DNS, and n/2 pairs of RNS (n=150000 in this study). Some training sample examples are shown in Figure 7.



(a) (b) (c)

Figure 7. Examples of training sample. The samples from top to bottom in (a) are a pair of positive sample, a pair of RNS, and a pair of DNS generated from visible-to-infrared images, respectively. (b) and (c) show samples generated from optical-to-SAR and optical-to-LiDAR images, respectively.

4.3.2 Network training: We trained FSNet on NVIDIA GTX 1080Ti GPU with Tensorflow. A batch size of 32 is used in each iteration and all the patches were resized to 97×97 pixels. The training is optimized by the Momentum Optimizer in Tensorflow. The initial learning rate and the momentum are set as 0.001 and 0.9 respectively. The training is terminated when the average loss value is less than 0.001. Figure 8 shows the convergence process of the training.

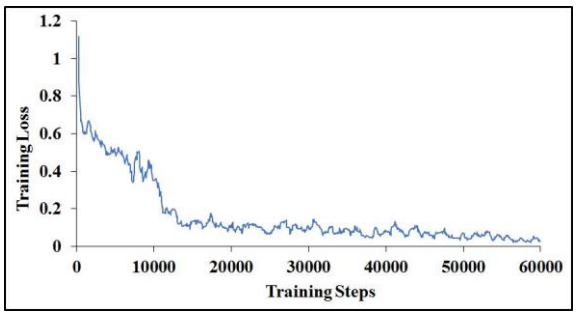


Figure 8. Training loss curve drew by Tensorboard.

4.4 Multimodal image matching results

The statistical results, including NCM and NTM, are displayed in Table 1. It can be seen directly from Table 1 that the proposed method performed best in term of NCM while the SIFT, CFOG and FSNet methods have not found any correct match on all image pairs. It is also indicated that the proposed method achieved the highest MP on the basis of the NCM and NTM values. The MP values are 100.00%, 93.67%, 79.88%, 63.19%, 76.92% and 59.76%, respectively.

Datasets	Statistical results (NCM/NTM)			
	SIFT	CFOG	FSNet	Proposed
pair 1	0/4	0/9	0/55	234/234
pair 2	0/5	0/8	0/58	222/237
pair 3	0/3	0/9	0/60	131/164
pair 4	0/2	0/9	0/59	103/163
pair 5	0/4	0/9	0/52	130/169
pair 6	0/4	0/8	0/53	98/164

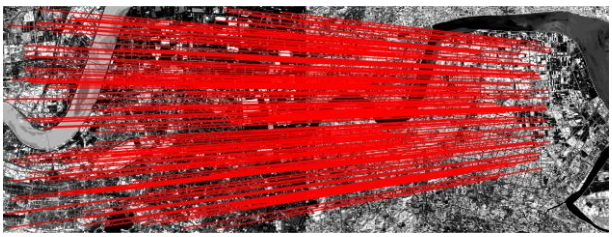
Table 1. Statistical matching results of the proposed method and all compared methods on all experimental datasets.

There is a normalization step in the descriptor computation in SIFT method, which makes it robust to illumination variation. However, the normalization is not robust to non-linear illumination difference. Therefore, SIFT method does not perform well for multimodal images with non-linear radiometric distortion although it is scale and rotation invariant.

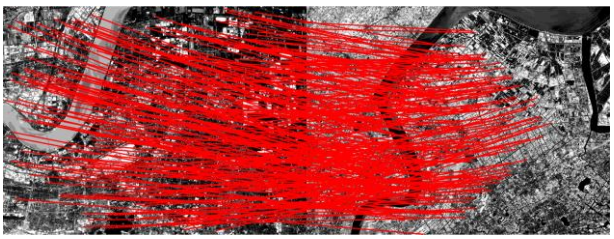
The CFOG method is designed for multimodal images matching especially. Some prior information, including direct geo-referencing information and GSD, must be available to correct the global geometric distortion between images to make the method work. If the prior information is unavailable, we have to select some GCPs manually to achieve coarse registration. However, in our experiments, we do not have any prior information of the images and enough GCPs to overcome the problem caused by geometric distortion. Therefore, the CFOG method failed in all image pairs with scale and rotation changes.

The FSNet method is a kind of deep learning-based method. Because geometric distortion has not been considered in the patch generation before similarity measurement, the patches of corresponding points are inconsistent. Thus, the features extracted from the patches are dissimilar and finally cannot be matched.

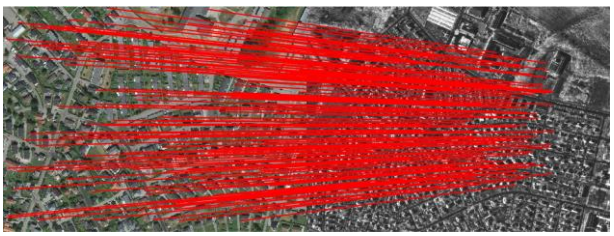
Compared with the aforementioned methods, both non-linear radiometric difference and geometric distortion are considered in the proposed method. On one hand, we exploit the deep feature maps to estimate the global geometric transformation between multimodal images and register images automatically. The coarse registration works without any extra information other than the two input images. After that, patches are produced from the coarsely registered images. Therefore, the proposed method is robust to image geometric distortion. On the other hand, we use a deep neural network trained by a dataset containing multimodal image patches with non-linear radiometric difference to extract features and measure similarity. The feature extraction and similarity measurement are robust to non-linear radiometric distortion. Therefore, the proposed method performs best on all image pairs. Figure 9 presents the matching results of the proposed method on all image pairs. Matches are linked with red lines.



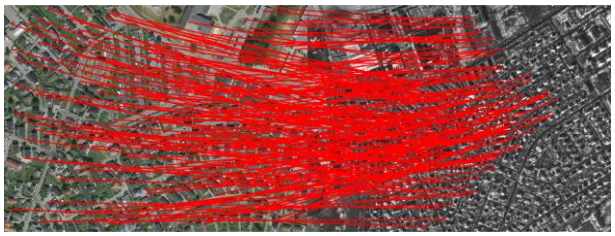
(a)



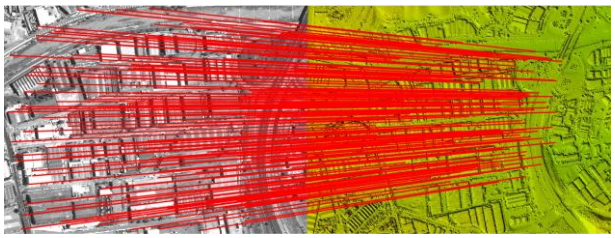
(b)



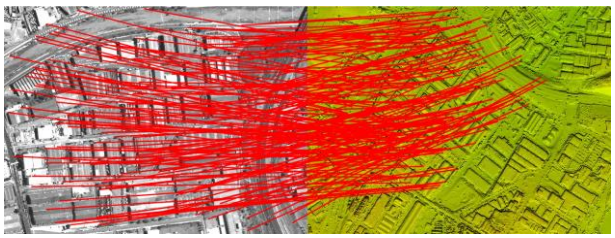
(c)



(d)



(e)



(f)

Figure 9. Matching results of the proposed method. (a)-(f) are the matching results of image pairs 1-6, respectively.

In addition, we can see from Table 1 that the proposed method performs better on image pairs 1 and 2 than on image pairs 3-6. There are two main reasons. First, the production of visible-to-infrared training samples is easier than that of optical-to-SAR and optical-to-LiDAR training samples. Therefore, the number of visible-to-infrared samples is larger than that of other two kinds of samples in our training datasets. It makes the trained neural network perform better on visible-to-infrared image pairs. Second, in the evaluation, it is easier to select accurate GCPs from visible-to-infrared image pairs than optical-to-SAR and optical-to-LiDAR image pairs. Therefore, the estimated image transformation to recognize correct and false matches is more accurate. In this case, some correct matches on optical-to-SAR and optical-to-LiDAR image pairs may be wrongly counted as false matches.

5. CONCLUSION

In this study, a multimodal image matching method that is robust to both geometric and non-linear radiometric distortion is proposed. We observed that the non-linear radiometric distortion between some deep feature maps generated from the last convolutional layer has been eliminated or relieved. On the basis of this observation, we analyzed the deep feature maps and designed a framework to overcome the geometric distortion between multimodal images. In this process, the geometric transformation can be estimated by using traditional feature matching method. We do not have to try to construct a feature descriptor that is robust to both geometric and non-linear

radiometric distortion under the help of deep feature maps. The experimental results demonstrate that the proposed method performs better than other state-of-the-art multimodal image matching methods. The proposed method can be used to match multimodal images without any prior information and manually selected GCPs. A possible future work is to increase training samples and make the proposed method work well on multimodal images in different areas.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No. 41971411 and 41631174) and the National Key Research and Development Program of China (No. 2016YFB0502603).

REFERENCES

Altwaijry H., Trulls E., Hays J., Fua P., Belongie S., 2016. Learning to match aerial images with deep attentive architecture. *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 3539-3547.

Bay H., Ess A., Tuytelaars T., Gool L., 2008. Speeded-up robust features (SURF). *Comput. Vis. Image Und*. 110(3), 346-359.

Chen H. M., Arora M. K., Varshney P. K., 2003. Mutual informationbased image registration for remote sensing data. *Int. J. Remote Sens*., 24(18), 3701-3706.

Chen M., Habib A., He H., Zhu Q., Zhang W., 2017. Robust Feature Matching Method for SAR and Optical Images by Using Gaussian-Gamma-Shaped Bi-Windows-Based Descriptor and Geometric Constraint. *Remote Sens*., 9(9), 882.

Gruen A., 2012. Development and Status of Image Matching in Photogrammetry. *Photogramm. Rec*., 27(137), 36-57.

Han J., Moraga C., 1995. The influence of the sigmoid function parameters on the speed of backpropagation learning. *in Proceedings of International Workshop on Artificial Neural Networks*, 195-201.

Harris C., Stephens M., 1988. A combined corner and edge detector. *in Proc. Alvey Vis. Conf.*, 147-152.

He H., Chen M., Chen T., Li D., 2018. Matching of remote sensing images with complex background variations via Siamese convolutional neural network. *Remote Sens*., 10(2), 355.

He H., Chen M., Chen T., Li D., Cheng P., 2019. Learning to match multitemporal optical satellite images using multi-support-patches Siamese networks. *Remote Sens. Lett*., 10(6), 516-525.

Hel-Or Y., Hel-Or H., David E., 2014. Matching by tone mapping: Photometric invariant template matching. *IEEE Trans. Pattern Anal. Mach. Intell*., 36(2), 317-330.

Huang J., You S., Zhao J.P., 2011. Multimodal Image Matching using Self Similarity. *in Proc. IEEE Applied Imagery Pattern Recognition Workshop*, 1-6.

Kong B., Supancic J., Ramanan D., Fowlkes C. C., 2019. Cross-Domain Image Matching with Deep Feature Maps. *Int. J. Comput. Vis*., 127: 1738-1750.

Kovesi P., 1999. Image Features from Phase Congruency. *Videre: J. Comput. Vis. Research*, 1(3), 1-26.

Lowe D. G., 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis*., 60(2), 91-110.

Melekhov I., Kannala J., RahtuE., 2017. Siamese network features for image matching. *in Proc. Int. Conf. Pattern Recognit*., 378-383.

Miller J.W., Goodman R., Smyth P., 1993. On loss functions which minimize to conditional expected values and posterior probabilities. *IEEE Trans. Inform. Theory*, 39(4), 1404-1408.

Morel J.M., Yu G., 2009. ASIFT: A New Framework for Fully Affine Invariant Image Comparison. *SIAM J. Imaging Sci.*, 2(2), 1-31.

Philbin J., Chum O., Isard M., Sivic J., Zisserman A., 2007. Object retrieval with large vocabularies and fast spatial matching. *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1-8.

Quan D., Liang X., Wang S., Wei S., Li Y., Huyan N., Jiao L., 2019. AFD-Net: Aggregated Feature Difference Learning for Cross-Spectral Image Patch Matching. *in Proc. IEEE Int. Conf. Comput. Vis.*, 3017-3026.

Rublee E., Rabaud V., Konolige K., Bradski G., 2011. ORB: An efficient alternative to SIFT or SURF. *in Proc. IEEE Int. Conf. Comput. Vis*., 2564-2571.

Sedaghat A., Mohammadi N., 2019. Illumination-Robust remote sensing image matching based on oriented self-similarity. *ISPRS J. Photogramm. Remote Sens*., 153, 21-35.

Ye Y., Bruzzone L., Shan J., Bovolo F., Zhu Q., 2019. Fast and Robust Matching for Multimodal Remote Sensing Image Registration. *IEEE Trans. Geosci. Remote Sens*., 57(11), 9059-9070.

Ye Y., Shen L., 2016. Hopc: A novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching. *in Proc. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci*., 9-16.

Zagoruyko S., Komodakis N., 2015. Learning to compare image patches via convolutional neural networks. *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 4353−4361.

Zhang J., Ma W., Wu Y., Jiao L., 2019. Multimodal remote sensing image registration based on image transfer and local features. *IEEE Geosci. Remote Sens. Lett*., 16(8), 1210-1214.