FAST AND ROBUST REGISTRATION OF MULTIMODAL REMOTE SENSING IMAGES VIA DENSE ORIENTATED GRADIENT FEATURE

Yuanxin YE [a,b]*

[a] State-province Joint Engineering Laboratory of Spatial Information Technology for High-speed Railway Safety,
Southwest Jiaotong University, 611756, China - yeyuanxin@home.swjtu.edu.cn
[b] Collaborative innovation center for rail transport safety, Ministry of Education, Southwest Jiaotong University,
611756, China - yeyuanxin@home.swjtu.edu.cn

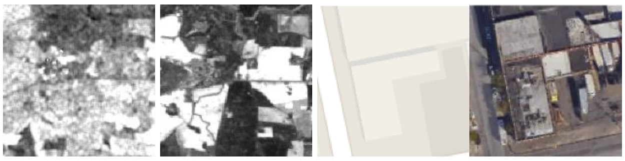
KEY WORDS: Multimodal Remote Sensing Image, Image Registration, DOGH, Similarity Metric

ABSTRACT:

This paper presents a fast and robust method for the registration of multimodal remote sensing data (e.g., optical, LiDAR, SAR and map). The proposed method is based on the hypothesis that structural similarity between images is preserved across different modalities. In the definition of the proposed method, we first develop a pixel-wise feature descriptor named Dense Orientated Gradient Histogram (DOGH), which can be computed effectively at every pixel and is robust to non-linear intensity differences between images. Then a fast similarity metric based on DOGH is built in frequency domain using the Fast Fourier Transform (FFT) technique. Finally, a template matching scheme is applied to detect tie points between images. Experimental results on different types of multimodal remote sensing images show that the proposed similarity metric has the superior matching performance and computational efficiency than the state-of-the-art methods. Moreover, based on the proposed similarity metric, we also design a fast and robust automatic registration system for multimodal images. This system has been evaluated using a pair of very large SAR and optical images (more than 20000×20000 pixels). Experimental results show that our system outperforms the two popular commercial software systems (i.e. ENVI and ERDAS) in both registration accuracy and computational efficiency.

1. INTRODUCTION

Image registration aims to align two or more images captured at different times, by different sensors or from different viewpoints (Zitova and Flusser 2003). It is a crucial step for many remote sensing image applications such as change detection, image fusion, and image mosaic. In the last decades, image registration techniques had a rapid development. However, it is still quite challenging to achieve automatic registration for multimodal remote sensing images (e.g., optical, SAR, LiDAR, and map), due to quite different intensity and texture patterns between such images. As shown in Figure 1, it is even difficult to detect correspondences (tie points) by visual inspection.



(a) (b)

Figure 1 Example of different intensity and texture patterns between multimodal remote sensing images. (a) SAR (left) and visible (right) images. (b) Map (left) and visible (right) images.

In general, image registration mainly includes three components (Brown 1992): feature space, similarity metric and geometric transformation. Feature space and similarity metric play the crucial roles in image registration.

The choice of feature space is closely related to image characteristics. A robust feature for multimodal registration should reflect the common properties between images, which are preserved across different modalities. Recently, Local invariant features such as Scale Invariant Feature Transform (SIFT) (Lowe 2004) and Speeded Up Robust Features (SURF) (Bay et al. 2008) have been widely applied to remote sensing image registration due to their robustness to geometric and illumination changes. However, these features cannot effectively detect tie points between multimodal images. This is because that they are sensitive to significant intensity differences, and cannot effectively capture the common properties between multimodal images (Suri and Reinartz 2010; Chen and Shao 2013).

Common similarity metrics include the sum of squared differences (SSD), the normalized cross correlation (NCC), the mutual information (MI), etc. These metrics are usually vulnerable to the registration of multimodal images because they are often computed using intensity information of images. In order to improve their robustness, some researchers applied these metrics on image features such as gradient and wavelet features. However, these features are not very effective for multimodal registration.

Recently, our researches show that structure and shape properties are preserved between different modalities (Ye and Shen 2016; Ye et al. 2017). Based on this hypothesis, tie points can be detected by using structure or shape similarity of images, which can be evaluated by calculating some traditional similarity metrics (e.g., SSD) on structure and shape descriptors. Additionally, the computer vision community usually uses pixel-wise descriptors to represent global structure and shape features of images, and such kind of feature representation has been successfully applied to object recognition (Lazebnik et al, 2006), motion estimation (Brox and Malik 2011), and scene alignment (Liu et al. 2011). Inspired from these developments,

* Corresponding author. Yuanxin YE, yeyuanxin@home.swjtu.edu.cn

we will explore the pixel-wise structural feature representation for multimodal registration.

In particularly, the contribution of this paper is that we first develop a pixel-wise feature descriptor that captures structure and shape features of images to address non-linear intensity and texture differences between multimodal images. This descriptor is named Dense Orientated Gradient Histogram (DOGH), which can be computed fast by convoluting orientated gradient channels with a Gaussian kernel. Then, a similarity metric based on DOGH is built in frequency domain, which is speeded up by Fast Fourier Transform (FFT), followed by a template matching scheme to detect tie points. Moreover, we also design a fast and robust automatic registration system based on DOGH for very large multimodal remote sensing images.

2. METHODOLOGY

Given a reference image and a sensed image, the aim of image registration is to find the optimal geometric transformation relationship between the two images. In practical application, we usually first detect the tie points between the images, and then use these tie pints to determine a geometric transformation model to align the images. In this section, we present a fast and robust registration method for multimodal remote sensing images, which includes the following aspects: (1) a per-pixel feature descriptor, named DOGH, is developed by using orientated gradients of images; (2) a similarity measure based on DOGH is proposed for tie point detection by a template matching scheme, and its computation is accelerated by FFT; (3) an automatic registration system is developed on the basis of CFOH and the proposed similarity measure , which can handle remote sensing images with the large size.

2.1 Dense Orientated Gradient Histogram

DOGH is inspired by Histogram of Orientated Gradient (HOG) (Dalal and Triggs 2005), which describes the shape and structural features by gradient amplitudes and orientation of images. HOG is calculated based on a dense grid of local histograms of gradient orientation over images, where the histograms are weighted by a trilinear interpolation method. Differently from that, DOGH is computed at every pixel of images based on local histograms of gradient orientation, and the histograms are quantized by applying a Gaussian filter in orientated gradient channels, instead of using the trilinear interpolation method. This will be much faster than HOG to compute the feature descriptor for every pixel of images.

We now give a formal definition of DOGH. For a give image, its M number of orientated gradient channels are first computed, which are referred as to g_i , $1 \le i \le M$. Each orientated gradient channel $g_o(x,y)$ equals the image gradient at location (x,y) for orientation o if it is larger than zero, else its value is zero. Formally, an orientated gradient channels is written as $g_o = \left\lfloor \frac{\partial I}{\partial o} \right\rfloor$, where I is the image, o is the orientation of the derivative, and $\lfloor \ \rfloor$ denotes that the enclosed quantity is equal to itself when its value is positive or zero otherwise. Then, each orientated gradient channel is convolved using a Gaussian kernel to achieve convolved feature channels as $g_o^\sigma = g_\sigma * \left\lfloor \frac{\partial I}{\partial o} \right\rfloor$, where σ is the value of Gaussian kernel. The final descriptor is 3D pixel-wise feature representation, which can capture the structural properties of images. Figure 2 shows the processing chain of DOGH,

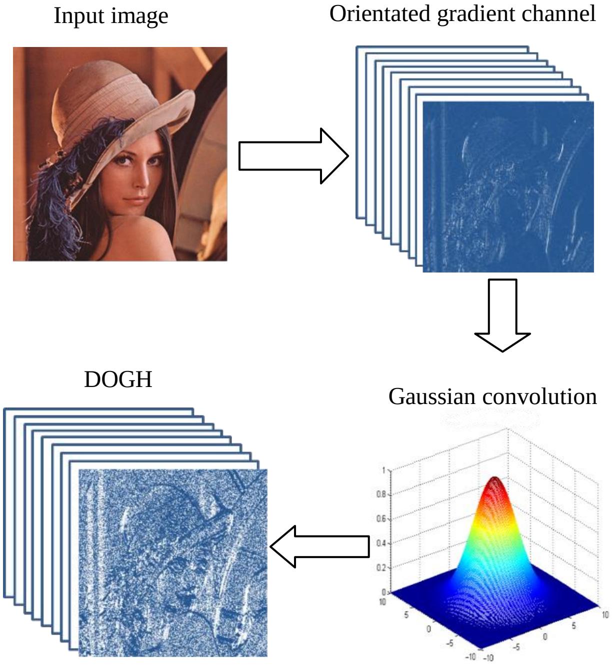


Figure 2 Processing chain of DOGH

2.2 Proposed similarity metric

This subsection proposes a similarity metric based on DOGH, and accelerate its computational efficiency by using FFT.

It is generally known that SSD is a popular similarity metric for image matching. For a reference image and a sensed image, let their corresponding DOGH be D_1 and D_2 , respectively. The SSD between the two DOGH can be computed by the following equation.

$$S_i(v) = \sum_x \left[D_1(x) - D_2(x\text{-}v) \right]^2 \tag{1}$$

where x is the location of a pixel in an image, and $S_i(v)$ denotes the SSD between D_1 and D_2 translated by a vector v over a template window i

In order to achieve the best match between D_1 and D_2 , it should minimize the similarity function $S_i(v)$. Accordingly, the matching function is

$$v_i = \arg\min_v \left\{ \sum_x \left[D_1(x) - D_2(x\text{-}v) \right]^2 \right\} \tag{2}$$

The obtained translation v_i is a translation vector that matches D_1 with D_2 for the template window i .

Since the pixel-wise structural feature representation is a 3D image which has a large data volume, it is time consuming to exhaustively compute the SSD similarity function for all candidate template windows. This is an intrinsic problem for template matching, as a template window needs to slide pixel-by-pixel within a search region for detecting its correspondences. An effective approach to reduce the computation of SSD is to use the FFT technique for acceleration.

At first, the matching function present in (1) is expanded as

$$S(v_i)=\sum_x D_1^2(x)+\sum_x D_2^2(x-v)-2\sum_x D_1(x)\bullet D_2(x-v) \quad (3)$$

In this equation, the first term are independent of v and can be efficiently calculated directly. The last two terms can be regarded as convolutions, which can be efficiently computed by using FFT since convolutions in the spatial domain become multiplications in the frequency domain. The final match function can be expressed as

$$v_i=\arg\min_v\left\{\begin{array}{l} C+F^{-1}\left[F(D_2)F^*(D_2)\right](v) \\ \quad -2F^{-1}\left[F(D_1)F^*(D_2)\right](v)\end{array}\right\} \quad (4)$$

where C represents the first term of the expanded form in (3), F and F^{-1} represents the forward and inverse FFTs, and F^* is the complex conjugate of F. This approach can reduce the computation of the similarity function. For example, given a template window with a size of $N\times N$ pixels and its search region is $M\times M$ pixels, the SSD takes $O(M^2N^2)$. Thus the proposed approach takes $O\left(\left(M+N\right)^2\log\left(M+N\right)\right)$ operations. Our approach has a substantial improvement in computational efficiency with the increase of the sizes of template window and search region.

2.3 Designed image registration system

Based on the proposed DOGH and similarity metric, we design an automatic image registration system for very large multimodal remote sensing images by using C++, which includes the following steps.

(1) The reference and sensed images are resampled to the same ground distance (GSD) to eliminate possible resolution differences.

(2) The block-based Harris operator (Ye and Shan 2014) is used to extract interest points in the reference image to make tie points distributed evenly over the image.

(3) After the extraction of interest points in the reference images, the proposed similarity metric based on DOGH (see section 2.2) is used to detect tie points in the sensed images by a template scheme, when the search region for image matching is predicted by the georeference information of images.

(4) Due to some factors such as occlusion and shadow, it is inevitable that obtained tie points have some errors. The tie points with large errors are removed using a global consistency check method based on a cubic polynomial model (Ma et al. 2010).

(5) After the removal of the tie points with large errors, a piecewise linear (PL) transformation model is applied to achieve image registration because this model can handle the local distortions caused by terrain relief to some degree.

3. EXPERIMENTS: DOGH MATCHING PERFORMANCE

In this section, DOGH is evaluated by using different types of multimodal remote sensing images. Two metrics such as the precision and computational efficiency are used to test the matching performance of DOGH. Moreover, DOGH is compared with the state-of-the-art similarity metrics such as NCC, MI and HOG$_{ncc}$ (Ye et al. 2017) to demonstrate its effectiveness. In the experiments, the parameter of DOGH is set to 9 orientation channels.

3.1 Data sets

We select a variety of multimodal images including visible, infrared, LiDAR, SAR and map data to test the proposed method. These data are divided into four categories: Visible-to-Infrared (Visib-Infra), LiDAR-to-Visible (LiDAR-Visib), Visible-to-SAR (Visib-SAR), and Image-to-Map (Img-Map). Before image matching, the reference and sensed images are resampled to the same ground sample distance (GSD) to remove possible differences in resolution. If images to be matched are the map data, they are rasterized. Figure 3 shows the test data, and Table 1 gives the description of these data.

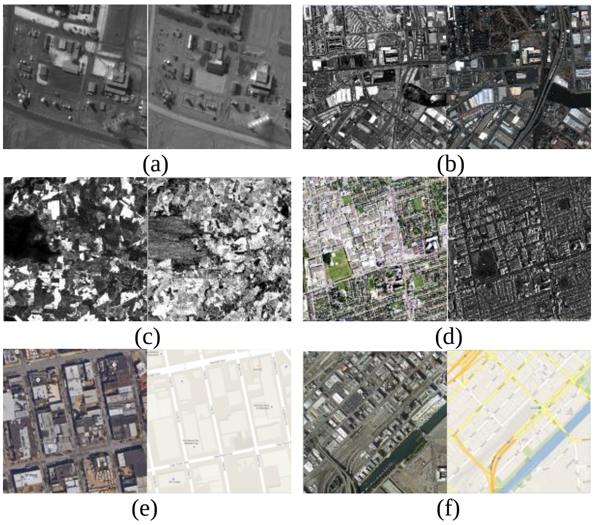

(a) (b) (c) (d) (e) (f)

Figure 3 Multimodal remote sensing images. (a) Test 1. (b) Test 2. (c) Test 3. (d) Test 4. (e) Test 5. (f) Test 6.

Category		Image pair	Size and GSD	Date
Visib-Infra	Test 1	Daedalus visible Daedalus infrared	512×512, 0.5m 512×512, 0.5m	2000/4 2000/4
LiDAR-Visib	Test 2	LiDAR intensity WorldView2 visible	600×600, 2m 600×600, 2m	2010/10 2011/10
Visib-SAR	Test 3	TM band3 TerraSAR-X	600×600, 30m 600×600, 30m	2007/5 2008/3
	Test 4	Google Earth TerraSAR-X	528×524, 3m 534×524, 3m	2007/11 2007/12
Img-Map	Test 5	Google Maps Google Maps	700×700, 0.5m 700×700, 0.5m	unknown
	Test 6	Google Maps Google Maps	621×614, 1.5m 621×614, 1.5m	unknown

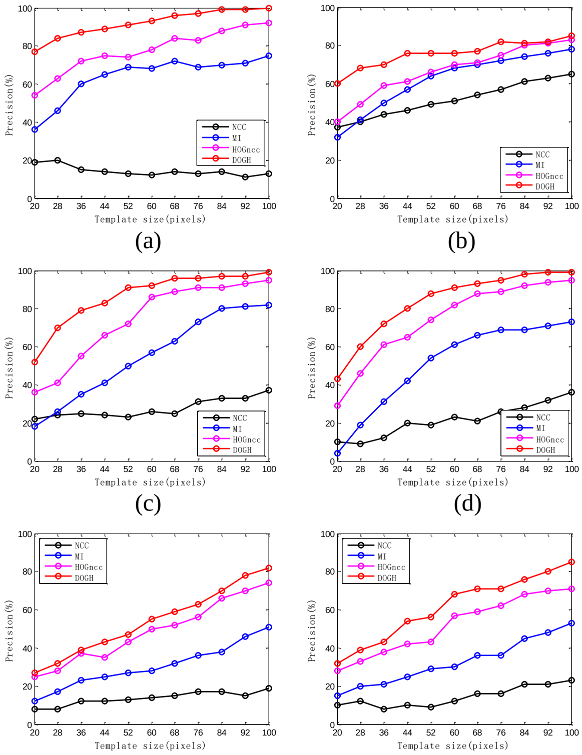
Table 1 Descriptions of the test data

3.2 Implementation details

The block-Harris detector is first used to extract the evenly distributed interesting points in the reference image. Then NCC, MI, HOG$_{ncc}$ and DOGH are used to detect the tie points in the sensed image by a template matching scheme, respectively. In order to analyze the sensitivities of similarity metrics with respect to changes in the template size, we use the template windows with different sizes to detect tie points between images.

3.3 Analysis of precision

The precision is defined as $presicion = CM / C$, where CM is the number of the correct match pairs, and C is the number of the total match pairs. Figure 4 shows the precision values of the four similarity metrics. It can be clearly observed that DOGH achieves the highest precision values in any template sizes. This shows that the similarity metrics representing structural similarity are more robust to complex intensity and pattern differences between multimodal images. NCC achieves the lowest precision values. This is because NCC is only invariant to linear intensity differences and is sensitive to complex intensity changes between images (Hel-Or et al. 2014). Although MI performs better than NCC, it still cannot effectively handle non-linear intensity differences between multimodal images.



(a) (b)

(c) (d)

(e) (f)

Figure 4 Precision values versus the template size of NCC, MI, HOG$_{ncc}$, and DOGH for multimodal images. (a) Test 1. (b) Test 2. (c) Test 3. (d) Test 4. (e) Test 5. (f) Test 6.

Compared with HOG$_{ncc}$, DOGH presents a superior matching performance. This is because DOGH depends on a dense (or pixel-wise) structural feature representation, which can better capture structural similarity between multimodal data than the relatively sparse feature representation used to construct HOG$_{ncc}$. Overall, the above experimental results demonstrate that DOGH is robust to the non-linear intensity difference between multimodal images.

3.4 Analysis of computational efficiency

Here, we analyze the computational efficiency of the four similarity metrics. Figure 5 shows that the run time taken from NCC, MI, HOG$_{ncc}$ and DOGH versus the template size. HOG$_{ncc}$ is calculated by the designed fast matching scheme (Ye et al. 2017). One can see that DOGH takes the least run time than the other similarity metrics. This is because DOGH is a low dimension of feature descriptor, and it accelerates the computation of similarity evaluation by using FFT.

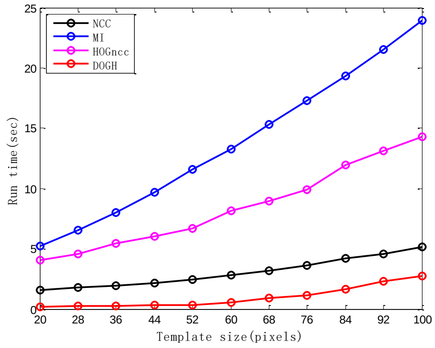


Figure 5 Run time versus the template size to NCC, MI, HOG$_{ncc}$, and DOGH.

4. EXPERIMENTS: MULTIMODAL REGISTRATION

To validate the effectiveness of our designed automatic registration system based on DOGH (see Section 2.3), two popular commercial software systems (i.e. ENVI 5.0 and ERDAS 2013) are used for comparison. Both ENVI 5.0 and ERDAS 2013 have the function modules for automatic remote sensing image registration, of which names are “Image Registration Workflow (ENVI)” and “AutoSync (ERDAS)”, respectively. Considering that the sizes of remote sensing images are usually large in the practical application, we use a pair of very large multimodal images (more than 20000×20000 pixels) for this comparison

4.1 Data sets

In the experiment, a pair of very large SAR and optical images is used to compare our systems with ENVI and ERADS. Figure 6 shows the test images, and Table 2 reports the description of the images. The challenges of registering the two images are as follows.

Geometric distortions: the images cover different terrains including mountain and plain areas. Different imaging modes between the SAR and optical images result in complex global and local geometric distortions between the two images.

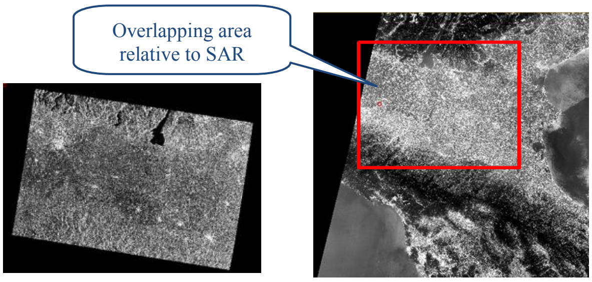
Intensity differences: significant non-linear intensity differences can be observed between the two images because they are captured by different sensors and at different spectral regions.

Temporal differences: the two images have a temporal difference of 12 months, which results in some ground objects

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W7-1009-2017 | © Authors 2017. CC BY 4.0 License.

changed.

Very large data volume: the SAR image and the optical image have the sizes of 29530×21621 pixels and 30978×30978 pixels, respectively.



(a) SAR image (b) Optical image

Figure 6 A pair of very large SAR and Optical images

Description	Reference image	Sensed image	Image characteristic
Sensor	Sentinel-1 SAR	Sentinel-2 Multispectral (optical) Instrument band 1	Images cover both mountain area and plain area. There is a temporal difference of 12 months between the images. Significant geometric and intensity differences. the SAR image is affected by significant noise
Resolution	10m	10m	
Date	07/2015	07/2016	
Size (pixels)	29530×21621	30978×30978	

Table 2 Descriptions of the test images

4.2 Implementation details

To our best knowledge, both ENVI and EARDAS apply the template matching scheme to detect tie points between images, which is the same as that of our systems. Accordingly, to make a fair comparison, all the systems set the same parameters for image matching, and use the PL transformation model for image correction. For the similarity metrics used in the ENVI and ERDAS, ENVI achieves image matching by NCC and MI, which are referred as “ENVI-NCC” and “ENVI-MI” in this paper, respectively. While ERDAS employs NCC to detect tie points, and uses a pyramid-based matching technique to enhance the robustness. Table 3 give the parameters used in all the systems. It should be note that because ERDAS uses a pyramid-based technique to guide the image matching, some parameters, such as search and template window sizes, cannot be set to too large. Therefore these parameters are set to the default values for ERDAS.

Parameter items	Our system	ENVI	ERDAS
Number of detected interest points	900	900	900
Search window size	80	80[1]	Default
Template window size	80	80	Default
Threshold value for error detection	3.5 pixels	3.5 pixels	Default

Table 3 Parameters used in all the systems.

4.3 Analysis of Registration Results

To evaluate the registration accuracy, we manually select 50 check points between reference and registered images, and employ the root mean-square error (RMSE) to represent the registration accuracy. Table 4 shows the registration results of all the systems. Our system outperforms the others, which includes achieving the most matched CPs, the least run time, and the highest registration accuracy.

Method	Tie points	Run time(sec)	RMSE(pixels)
Before-registration			18.65
ENVI-NCC	20	26.88	24.35(failed)
ENVI-MI	88	458.89	4.58
ERDAS	56	301.68	14.20
Our system	**303**	**19.24**	**2.33**

Table 4 Registration results of all the systems

ENVI-NCC fails in the image registration because its registration accuracy is worse than before-registration, while ENVI-MI and ERDAS improves registration accuracy compared with before-registration. For our system, it not only achieves higher registration accuracy than ENVI-MI and ERDAS, but also it is about 20x and 15x faster than ENVI-MI and ERDAS, respectively.

Figure 7 shows the registration results of before-registration, ENVI-MI, ERDAS, and our system. One can clearly see that our system performs best, followed by ENVI-MI and ERDAS.

The above experimental results show that our system is effective for the registration of very large multimodal images, and outperforms ENVI and ERDAS in both registration accuracy and computational efficiency.

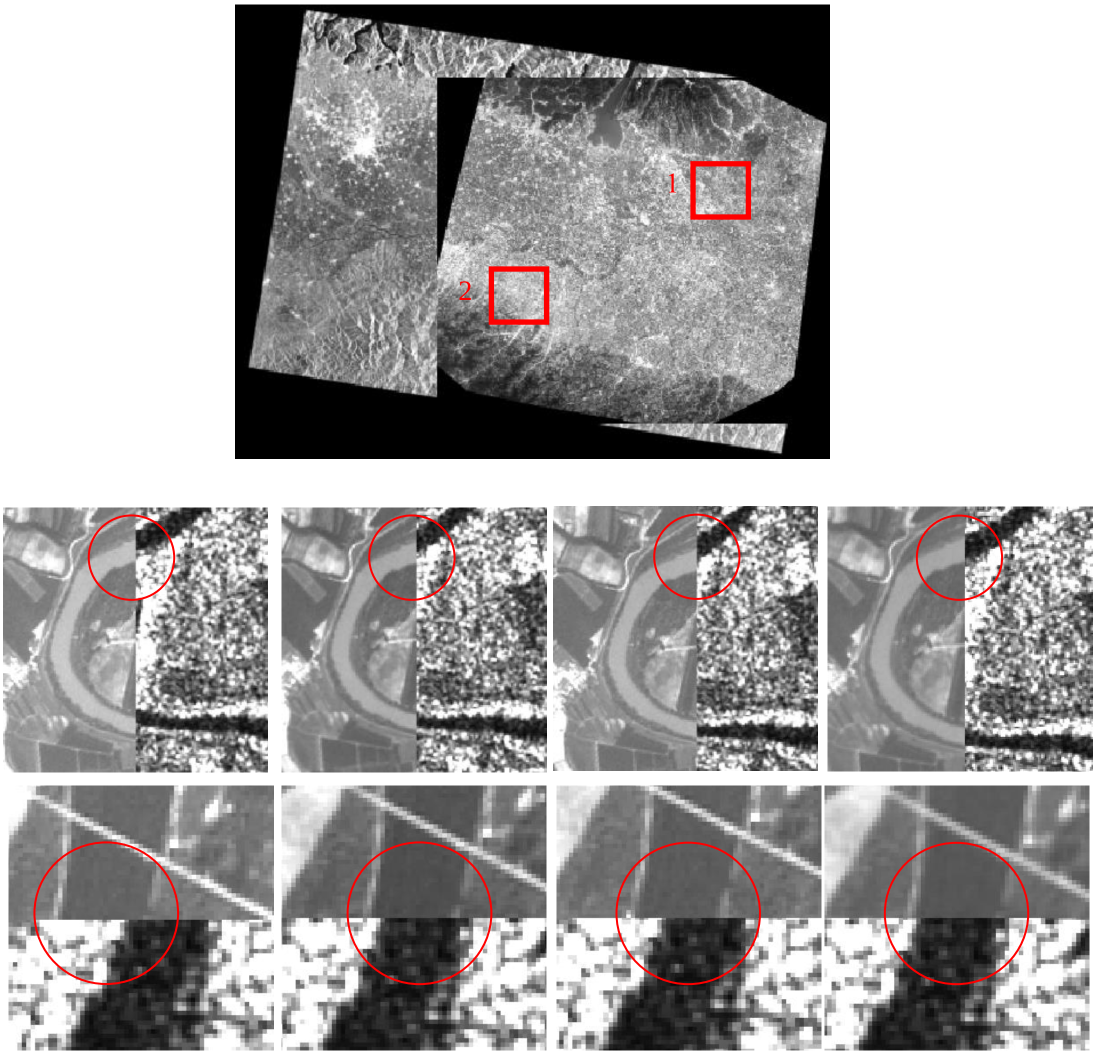
5. CONCLUSIONS

This paper proposes a fast and robust method for the registration of multimodal remote sensing images, to address non-linear intensity differences between such images. Our method is based on the proposed pixel-wise feature descriptor (named DOGH), which can capture structural properties of images. A fast similarity metric is designed for DOGH by FFT, which detects tie points between images using a template matching scheme. Six pairs of multimodal images are used to evaluate the proposed method. Experimental results show that DOGH performs better than the state-of-the-art similarity metrics such as HOG_{ncc}, MI and NCC.

In addition, an automatic images registration system is developed based on DOGH. The experimental results using a pair of very large SAR and optical images show that our system outperforms ENVI and ERDAS in both registration accuracy and computational efficiency. Especially for computational efficiency, our system is about 20x faster than ENVI, and 15x faster than ERDAS, respectively. This demonstrates that our system has the potential of engineering application. In apart from the registration of SAR and optical images, our system can also address the registration of other types of multimodal remote sensing data, such as optical, LiDAR and map. The more experiments will be present in future.

[1] Note: In the interface of ENVI, the “search window size” should be 160 pixels because it is equal to the sum of “search window size” and “template window size” in Table 3

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W7-1009-2017 | © Authors 2017. CC BY 4.0 License.



Before registration ENVI-MI ERDAS Our System

Figure 7 Registration results of before registration, ENVI-MI, ERDAS, and our system. Line 1 shows the registration results in the overlapping area of SAR and optical images. Line 2 shows the enlarged registration results in box 1. Line 3 shows the enlarged registration results in box 2.

ACKNOWLEDGEMENTS

This paper is supported by the National key Research Program of China (No.2016YFB0502603) and the National Natural Science Foundation of China (No.41401369).

REFERENCES

Zitova, B. and Flusser J., 2003. Image registration methods: a survey. *Image and Vision Computing*, 21(11), pp. 977-1000.

Brown, L. G., 1992. A survey of image registration techniques. *ACM computing surveys (CSUR)*, 24(4), pp. 325-376.

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

Bay, H., Ess, A., Tuytelaars,T., et al, 2008. Speeded-up robust features (SURF). *Comput. Vision Image Understanding.*, 110(3), pp. 346-359.

Suri, S. and Reinartz, P., 2010. Mutual-information-based registration of TerraSAR-X and ikonos imagery in urban areas. *IEEE Transactions on Geoscience and Remote Sensing*, 48(2), pp. 939-949.

Chen, M. and Shao, Z., Robust affine-invariant line matching for high resolution remote sensing images. Photogrammetric Engineering & Remote Sensing, 2013, 79(8), pp. 753-760.

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

Ye, Y., Shan, J., Bruzzone, L., et al, 2017. Robust registration of multimodal remote sensing images based on structural similarity. *IEEE Trans. Geosci. Remote Sens.*, 55(5), pp. 2941-2958.

Ye, Y., Shen, L., Hao, M. et al., 2017. Robust optical-to-SAR image matching based on shape properties, *IEEE Geosci. Remote Sens. Lett.*, 14(4), pp. 564-568.

Lazebnik, S., Schmid, C., and Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *Proc. CVPR2006*, pp. 2169-2178.

Brox, T. and Malik, J., 2011. Large displacement optical flow: descriptor matching in variational motion estimation, *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(3), pp. 500-513.

Liu, C., Yuen, J. and Torralba, A., 2011. Sift flow: Dense correspondence across scenes and its applications, *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5), pp. 978-994.

Dalal, N. and Triggs B., 2005. Histograms of oriented gradients for human detection. *Proc. CVPR2005*, pp. 886-893.

Cole-Rhodes, A. A., Johnson, K. L., LeMoigne, J., et al, 2003. Multiresolution registration of remote sensing imagery by optimization of mutual information using a stochastic gradient. *IEEE Transactions on Image Processing*, 12(12), pp. 1495-1511.

Dalal,N and Triggs B., 2005. Histograms of oriented gradients for human detection. *Proc. IEEE Conf. Computer Vision and Pattern Recognition 2005*, pp. 886-893

Ye, Y. and Shan J., 2014. A local descriptor based registration method for multispectral remote sensing images with non-linear intensity differences. *ISPRS Journal of Photogrammetry and Remote Sensing*, 90(2014), pp. 83-95.

Ma, J. L., Chan, J. C. W. and Canters, F., 2010. Fully automatic subpixel image registration of multiangle CHRIS/Proba data. *IEEE Transactions on Geoscience and Remote Sensing,* 48(7), pp. 2829-2839.

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

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W7-1009-2017 | © Authors 2017. CC BY 4.0 License.