# Multilevel SIFT Matching for Large-Size VHR Image Registration

Chunlei Huo, *Member, IEEE*, Chunhong Pan, Leigang Huo, and Zhixin Zhou

***Abstract*—A fast approach is proposed in this letter for large-size very high resolution image registration, which is accomplished based on coarse-to-fine strategy and blockwise scale-invariant feature transform (SIFT) matching. Coarse registration is implemented at low resolution level, which provides a geometric constraint. The constraint makes the blockwise SIFT matching possible and is helpful for getting more matched keypoints at the latter refined procedure. Refined registration is achieved by blockwise SIFT matching and global optimization on the whole matched keypoints based on iterative reweighted least squares. To improve the efficiency, blockwise SIFT matching is implemented in a parallel manner. Experiments demonstrate the effectiveness of the proposed approach.**

***Index Terms*—Coarse-to-fine strategy, geometric constraint, large-size image registration, parallel-based architecture.**

## I. Introduction

IMAGE registration is an old yet still hot topic, and it is involved widely in remote sensing. During the last decades, the spatial resolution increases significantly. Compared with low-to-moderate resolution images, higher registration accuracy is required for very high resolution (VHR) remote sensing applications such as change detection. However, VHR image registration is more difficult due to the following factors.

First, the difficulty is mainly caused by the complexity of VHR remote sensing images, i.e., high intraclass and low interclass variabilities [1]. Existed registration approaches can be generally categorized into two major categories: area-based and feature-based methods [2]. Compared with area-based methods, feature-based methods are recommended in remote sensing. Recently, local features such as scale-invariant feature transform (SIFT) [3] and speeded up robust features (SURF) [4] bring new potentials for feature-based remote sensing image registration due to the scale invariance of the detector and the distinctiveness of the descriptor. However, the high intraclass and low interclass variabilities make similar objects more ambiguous, so feature description and feature matching is more difficult. As a result, local features can be successfully applied for the registration of the visible images achieved by a hand-held high-definition digital camera, but when we adopt it to align remote sensing images, a lot of incorrect matches appear, and they are difficult to remove.

Second, the other outstanding property of VHR images is the overwhelming increase in image size, which results in the prohibitive memory requirement and computational complexity. For example, the QuickBird panchromatic image of Beijing consumes the storage of 1.5 GB. For another example, about 2000 SIFT keypoints are extracted from a $512 * 512$ image, each SIFT feature is of 132-D, even if SIFT features are encoded with the unsigned char type, the storage of 2000 SIFT features is $132 * 2000/1024/1024 = 0.25$ MB. For a $20\,000 \times 20\,000$ image, the storage of SIFT features is about $0.25 * 40 * 40 = 400$ MB. Due to the large size, the direct application of SIFT extraction and matching on VHR images is prohibitive on common desktop computers. The other impact caused by the image size increase is the ubiquitous repetitive structures (such as buildings and roads) represented in the VHR image, particularly in the urban scene, which makes the feature matching and outlier removal more challenging.

As for the first difficulty, many related approaches are reported in the literature [5], [6]. For example, Li *et al.* [5] proposed to use modified SIFT feature (i.e., feature descriptor refinement) and scale-orientation constraints to improve the matching performance, while Teke and Temizel [6] proposed to utilize the SURF feature and scale constraint for multispectral satellite image registration. The main difference between [5] and [6] lies in the difference between SIFT and SURF. In detail, SURF is superior to SIFT in efficiency, and SIFT is superior to SURF in case of scale, rotation, and blur. Local features of different types can achieve different performances [7], and it is important to compare them in the context of remote sensing image registration. However, compared to the first difficulty, the second one is less developed and more emergent for the practical application. To our best knowledge, most of SIFT-based approaches are designed for small-size images, and the approaches related to large-size image registration are rarely reported. For space limitation, we focus on addressing the second difficulty in this letter, i.e., a multilevel SIFT matching approach is proposed to address the difficulties caused by the overwhelming increase in image size. The rationale of the proposed approach is to reduce false matches caused by repetitive structures with the help of the geometrical constraint and to address the memory requirement and computational complexity by coarse-to-fine strategy and parallel blockwise local matching.

Manuscript received April 13, 2011; revised June 17, 2011 and June 27, 2011; accepted July 13, 2011. Date of publication September 8, 2011; date of current version February 8, 2012. This work was supported by the Natural Science Foundation of China under Grants 61005013, 60873161, and 60723005.

C. Huo and C. Pan are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: clhuo@nlpr.ia.ac.cn; chpan@nlpr.ia.ac.cn).

L. Huo is with the Department of Computational Mathematics, Xidian University, Xi'an 710071, China (e-mail: leiganghuo@163.com).

Z. Zhou is with the Beijing Institute of Remote Sensing, Beijing 100854, China (e-mail: zxzhou@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2011.2163491

## II. Proposed Approach

For large-size VHR image registration, the repetitive structure is the main reason that causes false matches. Recent work [8] indicates that the usage of priors is an efficient solution. However, this is not known before feature matching. To this end, we employ a coarse-to-fine strategy. Coarse registration is implemented at low resolution level based on the modified SIFT matching. With the help of the geometrical constraint and the transformation obtained at the coarse registration step, refined registration is implemented on the original image pairs by blockwise SIFT extraction and matching.

### A. Coarse Registration at Low Resolution Level

The first step of the proposed approach is to reduce the original image pairs to low resolution level, and this can be done by direct subsampling or pyramid decomposition (such as wavelet pyramid and Gaussian pyramid). Since SIFT will be re-extracted at high resolution level, considering the computation efficiency, direct subsampling is used in this letter. Due to the small size of low resolution images, SIFT extraction and matching can be implemented efficiently. To reduce the impacts of outliers, we use the traditional nearest distance ratio method [3] to get initial matched keypoints, scale-orientation joint restriction [5] is then used to reject false matches, and, finally, RANdom SAmple Consensus (RANSAC) is used to further remove outliers. The traditional SIFT-based registration uses a least squares method to solve the transformation parameters. However, it considers each matched pair equally in contribution to the solution of the final transformation, and this is too simple, particularly for large-size image registration. To estimate the transformation parameters more accurately, in this letter, we adopt an iterative reweighted least squares technique [9]. After estimating the transformation parameters at low resolution level, the approximated transformation at high resolution level can be obtained by modifying the shift parameters.

It is worth noting that the downsampling ratio is dependent on the spatial resolution and size of the input images. Similar to the technique to determine the wavelet decomposition level, the downsampling level $n$ is chosen as follows: $n = \lfloor \log 2(N/M) \rfloor$, and the downsampling ratio $r = 2^n$, where $\lfloor \cdot \rfloor$ is the floor function, $N$ is the minimum of the width and height of two input images, and $M$ is the user-defined minimized size of the image after downsampling. For large-size images, $M$ cannot be too small. With the downsampling ratio increase, more details are reduced, and SIFT matching may fail when the downsampling ratio is too large. In general, $r$ is equal to or smaller than 16. For example, for a $20\,000 \times 20\,000$ image pair, $r = 8$ is a good choice.

### B. Refined Registration at High Resolution Level

With the help of the transformation parameters achieved at low resolution level, we can apply blockwise SIFT extraction and matching to improve the efficiency. False matches caused by the repetitive structures are difficult to remove if keypoints are matched purely based on SIFT descriptor similarity. While with the help of the geometrical constraint provided by the approximated transformation, the proposed approach can reject such false matches easily, this is different from and superior to the traditional global matching (this topic will be discussed in detail in the experiment section). By collecting the matched keypoints from each block pair, we can get a large matched keypoint set. Then, the residual error based on the initial transformation is used to remove outliers. Finally, accurate transformation parameters can be estimated by the global optimization on the whole matched keypoints based on iterative reweighted least squares.

For large-size image registration, the main challenges are the prohibitive computation complexity and overwhelming memory requirement. The challenges described earlier lie in the following two facts. First, direct SIFT extraction on large-size images is impossible on a desktop computer. A feasible approach is to split the large-size image into blocks and apply SIFT on the individual block. However, due to the dense nature of SIFT, the storage of such a SIFT set is difficult. Second, even if the computation time is permitted, the traditional global matching of such two huge size SIFT sets is not reliable due to the impacts caused by the repetitive structures, while with the help of the geometrical constraint, blockwise SIFT extraction and matching makes it possible. Based on the proposed blockwise strategy, SIFT extraction and matching is implemented on two coarsely aligned image blocks, and only the matched SIFT features are saved, so there is no problem in applying SIFT extraction and matching on large-size images even on a personal desktop computer. Aside from the efficiency and robustness, blockwise matching is helpful to make the matched keypoints well distributed, which is important to improve the registration accuracy.

### C. Parallel Architecture

To further improve the efficiency, we implement the blockwise SIFT extraction and matching procedure in a parallel fashion, i.e., we construct a high-performance platform by a cluster of multikernel PCs connected by the very high-speed network, and we adopt a master–slave model to distribute the image data to the cluster system. The PC cluster consists of one master node and some slave nodes. The master node controls the whole cluster and sends tasks to the slave nodes. For our algorithm, the task means blockwise SIFT extraction and matching. In detail, after finishing coarse registration at the master node, regular image block pairs are sent to the slave nodes for blockwise SIFT extraction and matching. The minimum computing unit of the slave node is the kernel. Once a task is finished by a slave node, the result is sent to the master node. By this way, the master node monitors the resource and assigns the task to the idle CPU. The master node merges the matched keypoints after all tasks are finished. Then, outlier removal, transformation estimation, image resampling, and transformation are implemented at the master node. The flowchart of the proposed approach is illustrated in Fig. 1.

## III. Experiments and Discussion

To assess the effectiveness of the proposed approach in detail, two sets of experiments are designed. The first set of experiments is to illustrate the importance of the geometrical constraint, and the second set of experiments aims to demonstrate the advantages of coarse-to-fine strategy, blockwise matching,

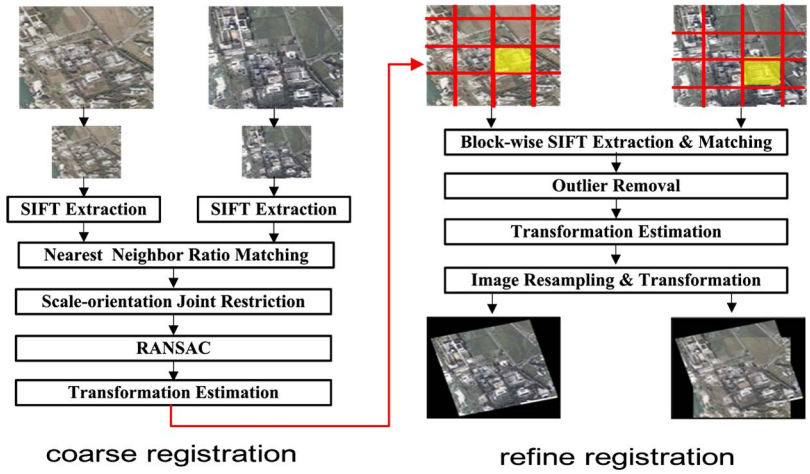


Fig. 1. Flowchart of the proposed approach.

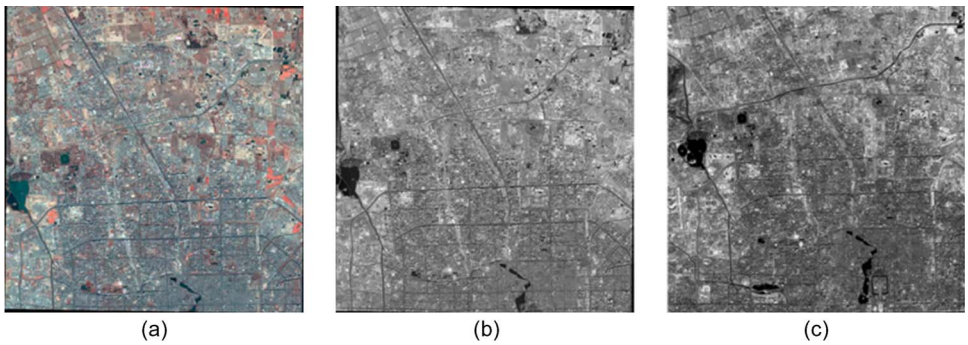


Fig. 2. Images used in this letter. (a) Multispectral image of 2002. (b) Panchromatic image of 2002. (c) Panchromatic image of 2003.

and parallel architecture. The images are taken over Beijing (China) acquired by the QuickBird satellite. As shown in Fig. 2, the images consist of a multispectral image (April, 9, 2002, 2.4 m/pixel, 8986 × 8504) and panchromatic images (one is on April, 9, 2002, 0.61 m/pixel, 29 014 × 27 552; the other is on November 12, 2003, 0.63 m/pixel, 30 969 × 27 610).

### A. Experiment

To illustrate the importance of the geometrical constraint, we compared the proposed approach (local matching) with the traditional global matching (initial matching based on SIFT descriptor similarity + scale-orientation joint constraints + RANSAC). In this experiment, we use the following measures to evaluate the effectiveness of the geometrical constraint: the number of matches, the number of correct matches, and the matching ratio. The images used in this experiment are shown in Fig. 3, which are subimages of Fig. 2. The performances of different approaches are listed in Table I.

For data set 1 [see Fig. 3(a)], two images are similar in the spatial resolution. In this case, 3706 SIFT keypoints are matched based on the similarity of the SIFT descriptor. However, due to the inherent nature and the repetitive structure of VHR remote sensing images, the traditional matching technique based purely on the descriptor similarity is too fragile. As shown in Fig. 3(a), many initial matches are false. Furthermore, even with the help of scale-orientation constraints, such false matches are difficult to remove since they have

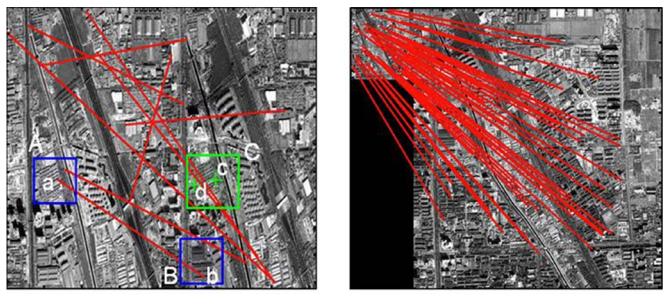


Fig. 3. Comparison of different approaches. (a) Some false matches based on global matching without geometrical constraint. (b) Some correct matches based on the proposed approach.

TABLE I
PERFORMANCE COMPARISON ON SIFT MATCHING

| | | global matching | local matching |
|---|---|---|---|
| dataset 1 | initial matches | 3706 | 1139 |
| | correct matches | 354 | 672 |
| | matching ratio(%) | 9.6 | 59.0 |
| dataset 2 | initial matches | 977 | 1876 |
| | correct matches | 77 | 721 |
| | matching ratio(%) | 7.9 | 38.4 |

similar scale and orientation. As a result, only 359 pairs are achieved after RANSAC, among which 354 pairs are correct. In contrary, by the geometrical constraint, 1139 pairs are obtained based on SIFT descriptor similarity, and 672 correct matches are achieved. With the help of the geometrical constraint,

the matching ratio increases from 9.6% to 59.0%, and this demonstrates the advantages of the geometrical constraint in improving the performance. In detail, this improvement can be explained from the following two aspects: 1) The geometrical constraint is helpful for reducing the error propagation domain. For example, as illustrated in Fig. 3(a), by the traditional global approach, $a$ is wrongly matched to $b$, while for the proposed approach, the false matching domain $B$ is excluded, and the matching domain of the keypoints in the image block $A$ is confined to the image block $C$. In other words, matching is of propagation, and by the geometrical constraint, the error propagation domain is reduced from the whole image to the specified local block; this will eliminate false matches significantly. 2) The geometrical constraint is beneficial for validating the candidate matches. Due to the ubiquitous repetitive structures, even by blockwise matching, many false matches cannot be removed without the help of the geometrical constraint. For instance, if purely based on the SIFT descriptor, $a$ in image block $A$ is similar to two keypoints in image block $C$, i.e., $c$ and $d$. Furthermore, due to the impacts caused by the disturbances such as view angle, occlusion, and noise, the similarity between $a$ and $d$ may be greater than that between $a$ and $c$, and the traditional nearest neighbor matching will fail, while this situation will change if the geometrical constraint is under consideration. In detail, supposing that the transformation achieved at the initial registration step is $T$, if the residual error between the transformed coordinate of $a$ and its candidate matched keypoints $p$ is larger than a threshold, i.e., $dis(T(a), p) \geq \tau$, then such matches are rejected. By this way, false matches between $a$ and $d$ are removed successfully. In other words, the pure usage of the SIFT descriptor for matching is not reliable, and the geometrical constraint is good for validating the candidate matches synergically. Since iterative reweighted least squares will be used to solve the optimal transformation parameters (where the threshold will be tuned adaptively), it is not necessary to choose the threshold $\tau$ very carefully. In this letter, $\tau = 100$.

For data set 2 [see Fig. 3(b)], the spatial resolution difference is about 4. In this case, feature matching is more difficult. The global matching can achieve 977 pairs, and 77 pairs are correct, while the proposed approach can achieve 1876 pairs and 721 pairs are correct. The matching ratio increases from 7.9% to 38.4%. The aforementioned comparisons indicate the advantage of the geometrical constraint in filtering out the false matches. Of course, such advantage is contributed to the coarse-to-fine matching strategy.

### B. Experiment

This experiment is implemented on three large-size VHR images shown in Fig. 2. To illustrate the efficiency of blockwise SIFT extraction/matching and parallel fashion, in this letter, we compared the proposed approach (the parallel approach) with the serial approach (blockwise SIFT extraction and matching on a desktop computer) in efficiency, i.e., the CPU time on feature extraction/matching (including the coarse step and the refined step). It is worth noting that the overall performances of the aforementioned two approaches are relative to the number of tasks and the computing environment (such as the performance of the PC, the number of PCs, and the bandwidth of the network). In this letter, our parallel environment consists of 11 desktop computers (Intel Core i7 920 Quad-Core CPU, 2.67 GHz, 4-GB DDR RAM). The algorithms are implemented in C++ and compiled under Microsoft Visual C++ 2005. MPICH2 [10] is used for parallel programming, including task monitoring and multiple kernel controlling.

TABLE II
PERFORMANCE COMPARISON AT CPU TIME

| | | serial approach | parallel approach |
|---|---|---|---|
| dataset 1 | CPU time(s) | 9230 | 972 |
| dataset 2 | CPU time(s) | 5013 | 784 |

Given the computing environment and the sizes of the input images, the performance is mainly determined by the block size, and this relation can be measured quantitatively. By experiments, we find that, for the computing environment described earlier, the best performances are achieved when the block size is $1024 \times 1024$ if the image size is larger than $4096 \times 4096$, and this conclusion holds for the parallel approach and the serial approach. For this reason, the following experiments are evaluated by setting the block size to be $1024 \times 1024$.

The performances of different approaches are listed in Table II. In this experiment, data set 1 means panchromatic images of 2002 [see Fig. 2(b)] and 2003 [see Fig. 2(c)], and data set 2 means the multispectral image of 2002 [see Fig. 2(a)] and panchromatic images of 2003 [see Fig. 2(c)]. From Table II, we can see that the proposed approach is superior to the serial approach in efficiency, and the CPU time is reduced about 9.5 and 6.4 times, respectively. This comparison demonstrates the advantages of parallel architecture in improving the efficiency.

To assess the proposed approach quantitatively, the mean distance between the matched SIFT pairs under the transformation computed on the real control point pairs provided by the bureau of surveying and mapping (i.e., rmse between the matched keypoints) is used to evaluate the registration accuracy. The mean distance between the matched SIFT set is 0.31 for data set 1 and 0.48 for data set 2. Noting the large size of the images being considered, such accuracy is promising. To evaluate the accuracy of the proposed approach visually, we merge the registered multispectral and panchromatic images in 2002 and the reference image in 2003; the results are shown in Fig. 4. Fig. 4(a) is the merged pseudocolor image of the whole scene, Fig. 4(b) is the checkerboard image of the multispectral image in 2002 and the panchromatic image in 2003, and Fig. 4(c) is the checkerboard image of the panchromatic images in 2002 and 2003. From Fig. 4, we can see that the transformed images are well matched to the reference image, and this demonstrates the effectiveness of the proposed approach.

As a final remark, it is worth noting that the transformation model is very important for image registration. For images captured by satellite imaging systems with narrow angular field of view over relatively flat terrain (a terrain with negligible height variations compared with the flying height), affine transformation is usually used in the literature. In consequence, for the images being considered earlier, affine transformation model is used at the coarse and refined steps. In addition, we compared affine and projective models and found that there is no significant difference. However, for aerial remote sensing images, a more advanced model should be considered at the refined step (since the role of the coarse registration step is to


Fig. 4. Image registration results. (a) Pseudocolor image of the whole scene. (b) Local checkerboard image of multispectral image in 2002 and panchromatic image in 2003. (c) Local checkerboard image of panchromatic image in 2002 and panchromatic image in 2003.


Fig. 5. Image registration on aerial images. (a) Aerial image in 2001. (b) Aerial image in 2002. (c) Merged pseudocolor image.

provide an initial transformation for the refined step and establish the block correspondence for blockwise SIFT extraction and matching and since SIFT will be re-extracted at the refined step, affine model is enough at the coarse registration step even for aerial image registration). To clarify this point, we compared affine, projective, and quadratic polynomial models on a pair of aerial images (see Fig. 5) and found that the best performance is achieved by the quadratic polynomial model $(\text{rmse} = 0.51)$ and the affine and projective ones are not adequate for the complex aerial images (rmse values are 0.62 and 0.57, respectively). For space limitation, automatic model selection is beyond the scope of this letter.

## IV. Conclusion

A fast approach is proposed in this letter for the registration of large-size VHR images. The main contributions of the proposed approach lie in the coarse-to-fine matching strategy and the parallel implementation of blockwise feature extraction/matching. The parallel architecture is benefited from the coarse registration at low resolution level, which provides the geometrical constraint for high resolution level and makes the blockwise SIFT feature extraction/matching possible. Despite the effectiveness and efficiency of the proposed approach, many developments need to be considered in the future work, including the algorithm optimization and the generation of the proposed approach for multisensor remote sensing image registration.

## Acknowledgment

The authors would like to thank RSIU-LIAMA (http://www.kesala.net/?page_id=36) for providing the QuickBird images and the anonymous reviewers for their suggestions to improve this letter.

## References

[1] A. Carleer, O. Debeir, and E. Wolffa, "Comparison of very high spatial resolution satellite image segmentation," *Proc. SPIE*, vol. 5238, pp. 532–542, 2004.
[2] B. Zitova, "Image registration methods: A survey," *Image Vis. Comput.*, vol. 21, no. 11, pp. 977–1000, Oct. 2003.
[3] D. Lowe, "Distinctive image features from scale invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[4] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.
[5] Q. Li, G. Wang, J. Liu, and S. Chen, "Robust scale-invariant feature matching for remote sensing image registration," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 2, pp. 287–291, Apr. 2009.
[6] M. Teke and A. Temizel, "Multi-spectral satellite image registration using scale-restricted SURF," in *Proc. ICPR*, 2010, pp. 2310–2313.
[7] K. Mikolajczyk and C. Schmid, "Comparison of affine-invariant local detectors and descriptors," in *Proc. Eur. Signal Process. Conf.*, 2004, pp. 1729–1732.
[8] H. Wang, S. Yan, J. Liu, T. Huang, and X. Tang, "Correspondence propagation with weak priors," *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 140–150, Jan. 2009.
[9] P. Meer, "Robust techniques for computer vision," in *Emerging Topics in Computer Vision*, G. Medioni and S. B. Kang, Eds. Englewood Cliffs, NJ: Prentice-Hall, 2004.
[10] MPICH2, 2010. [Online]. Available: http://www.mcs.anl.gov/research/projects/mpich2