An Efficient, SIFT-Based Mode-Seeking Algorithm for Sub-Pixel Registration of Remotely Sensed Images

Benny Kupfer, Nathan S. Netanyahu *Member, IEEE*, and Ilan Shimshoni *Member, IEEE*

***Abstract*—Several image registration methods, based on the scaled-invariant feature transform (SIFT) technique, have appeared recently in the remote sensing literature. All of these methods attempt to overcome problems encountered by SIFT in multi-modal remotely sensed imagery, in terms of the quality of its feature correspondences. The deterministic method presented in this paper exploits the fact that each SIFT feature is associated with a scale, orientation, and position to perform mode seeking (in transformation space) to eliminate outlying corresponding key points (i.e., features) and improve the overall match obtained. We also present an exhaustive empirical study on a variety of test cases, which demonstrates that our method is highly accurate and rather fast. The algorithm is capable of automatically detecting whether it succeeded or failed.**

***Index Terms*—Remotely sensed images, image registration, feature correspondence, mode-seeking SIFT.**

I. Introduction

IMAGE registration (IR) of multi-temporal, multi-spectral, and multi-sensor images is a fundamental building block in a variety of remote sensing applications, e.g., change detection, image mosiacing and classification, environmental monitoring, etc. Although many approaches have been proposed over the years for IR of remotely sensed data, the problem remains challenging to this day, and it requires various ongoing innovations, in an attempt to obtain enhanced performance, vis-à-vis accuracy, running time, etc. In this paper we propose an efficient IR method based on the *scale invariant feature transform* (SIFT) [1]. The feature invariance captured through descriptors based on image gradients makes SIFT features applicable to IR of image pairs acquired at different times, from different spectral bands, or by different sensors. This capability of our method is demonstrated empirically.

The method performs reliable filtering of outlying feature correspondences by *mode seeking* of scale ratios, rotation differences, and eventually horizontal and vertical shifts between all corresponding SIFT key points. This is done using the scale, orientation, and position associated with each key-point. Our deterministic *mode-seeking SIFT* (MS-SIFT) algorithm is very simple and fast and appears to achieve sub-pixel accuracy.

The paper is organized as follows. In Section II we provide a brief survey of related, SIFT-based methods. In Section III we describe our method, and in Section IV we present the main results of our extended empirical studies. Section V analyzes failed registrations and Section VI offers possible remedies for performance enhancement. Finally, Section VII provides concluding remarks.

II. Prior Work

In previous work [2], [3] we used edge-like wavelet features to perform feature matching by hierarchical searching in transformation space. An initial bounding box was used to reduce the elaborate search required at the higher resolution levels and the *partial Hausdorff distance* (PHD) was chosen as a similarity measure. Our new method is designed to alleviate, to a significant extent, the need for the above described search process. Other researchers also draw strongly on the notion of SIFT for image registration of remotely sensed data. However, since SIFT often results in inaccurate (if not incorrect) matching when applied to multi-modal remotely sensed imagery, the main thrust is to obtain first a reliable set of corresponding key points.

Li *et al.* [4] proposed to refine the SIFT key-point orientations and assign multiple orientations to each key-point. They perform outlier filtering based on the ratio between the Euclidean distance to the closest neighbor and the second closest neighbor (as proposed in [1]), followed by similar pruning with respect to the so-called *joint distance* (JD). An iterative search between all orientation differences is used to find the best match (in a JD sense) for the resulting key points. Teke *et al.* [5] proposed *orientation restricted SIFT* (OR-SIFT). Orientations with opposite directions are binned together to compensate for inversions in the gradient orientations, and matches are based on nearest-neighbor (NN) distances between SIFT features, where false matches are excluded if their SIFT key points scale distance is larger than a predefined threshold. Sedagat *et al.* [6] proposed the *uniform robust SIFT* (UR-SIFT) algorithm, where extracted SIFT key points are distributed evenly in both the scale and image spaces. Key-points with low principal curvature are rejected; further rejection is obtained by checking each correspondence in a global transformation model between the reference and sensed images. Li *et al.* [7] performed matching using the rotation-invariant distance between SIFT key points in a polar grid. Matching is done via RANSAC [8], followed by a final transformation computation. Hasan *et al.* [9] proposed a 2-step procedure which also rejects outliers according to the distance ratio between the first and second NNs. RANSAC is used to exclude remaining outliers, and a global transformation is computed according to so-called primary and secondary matched feature points. Finally, Hasan *et al.* [10] proposed

B. Kupfer is with the Department of Mathematics, Bar-Ilan University, Ramat-Gan 52900, Israel

N. S. Netanyahu is with the Department of Computer Science, Bar-Ilan University, Ramat-Gan 52900, Israel and with the Center for Automation Research, University of Maryland at College Park, MD 20742

I. Shimshoni is with the Department of Information Systems, Haifa University, Haifa 31905, Israel

A preliminary version of this paper appeared in *Proc. of the IEEE Geoscience and Remote Sensing Symposium*, 2013.

numerous modifications in the SIFT procedure, e.g., preserving every SIFT key-point, limiting gradient values to reduce the effect of strong edges, using a larger window for the SIFT descriptor, etc.

III. Proposed Method

The above described methods require typically thousands of SIFT key points in both the reference and sensed images, even for a standard image size. Also, some of them require exhaustive search and matching. Registration run-times are rarely reported, and those reported vary from tens to hundreds of seconds (for complete registration). In contrast, our method uses a relatively high threshold to detect initially only (up to) hundreds of SIFT key points for each image. Key points of the reference and sensed images are then matched according to NNs of corresponding SIFT descriptors. We abandon the traditional approach of filtering outliers according to the distance ratio between the first and second NNs [1]. This is done since even though this method eliminates a large number of outliers, it also eliminates quite a few inliers as well. RANSAC is also not used although it produces more accurate initial transformations, its complexity is much higher.

Instead, much in the spirit of a Hough-like voting scheme, we exploit the inherent information of each SIFT key-point (i.e., scale, orientation, and position) to compute a prospective transformation for each match (i.e., corresponding key points). We assume a similarity transformation model, implied naturally by the SIFT characteristics (i.e., scale, rotation, and vertical and horizontal translations); this model is widely used in IR of remotely sensed data as satellite images are usually nadir images, the scales at both axes are identical, and shear is practically insignificant. In principle, we perform mode seeking in 4-D space, which is done in practice for each of the four components separately. This is followed by effective pruning of outlying correspondences and a refined computation of the transformation.

We start by computing a histogram of the scale ratios for all of the SIFT key-point matches and find its mode scale, s_{mode}. Similarly, we compute a histogram of the orientation differences for all the matches and find its mode rotation difference, $\Delta\theta_{\text{mode}}$. Our experimental results show that for a variety of multi-temporal and multi-spectral images the histogram modes are unique and evident (at least 40% higher than the next peak). Even though each SIFT match has its own scale and orientation difference, we use the modes obtained from all the features since they are more accurate. Moreover, a mode of a distribution can be estimated even when there exist a large number of outliers. We therefore use the scale ratio and rotation difference modes found to perform mode seeking of the horizontal and vertical translations as follows. Let (x, y) and (x', y') denote, respectively, the coordinates of a SIFT key-point in the reference image and its corresponding key point in the sensed image. Each pair of corresponding key points defines the following horizontal and vertical shifts:

$$\Delta x = x - s_{\text{mode}}(x' \cos(\Delta\theta_{\text{mode}}) - y' \sin(\Delta\theta_{\text{mode}})), \tag{1}$$

$$\Delta y = y - s_{\text{mode}}(x' \sin(\Delta\theta_{\text{mode}}) + y' \cos(\Delta\theta_{\text{mode}})). \tag{2}$$

We now compute two additional histograms of Δx and Δy for all corresponding key points, for which we find the mode values, Δx_{mode} and Δy_{mode} , respectively. The quadruple obtained, $< s_{\text{mode}}, \Delta\theta_{\text{mode}}, \Delta x_{\text{mode}}, \Delta y_{\text{mode}} >$, is used (as a transformation approximation) to eliminate outlying key-point pairs according to the following logical filters:

$$|\Delta x - \Delta x_{\text{mode}}| \geq \Delta x_{\text{thresh}}, \tag{3}$$

$$|\Delta y - \Delta y_{\text{mode}}| \geq \Delta y_{\text{thresh}}, \tag{4}$$

where Δx and Δy are given in (1) and (2), and $\Delta x_{\text{thresh}}, \Delta y_{\text{thresh}}$ denote, respectively, thresholds of horizontal and vertical differences, in terms of corresponding histogram bin widths (measured in pixels). All corresponding pairs $(x, y) \Leftrightarrow (x', y')$ for which (3) or (4) hold will be considered outliers and thus rejected. Although this filters out typically 80%–90% of the initial correspondences (for the data sets we have experimented with), the resulting set of correspondences is very reliable, so that it suffices to employ at this stage a 1-step *ordinary least squares* (OLS) procedure. In a nutshell, this is done by first computing the transformation that aligns the centroids of the (remaining) point sets, then computing the scale factor that aligns their spatial variances, and finally computing the rotation that minimizes the sum of squared distances [3].

IV. Experimental Results

We have implemented our MS-SIFT procedure in C using a single thread style, and tested it on scores of remotely sensed image pairs [11]. Performance (in terms of accuracy) was evaluated in each case via manual ground truth (GT), using the *root mean square error* (RMSE) criterion. Picking manually N corresponding points $(x_i, y_i) \Leftrightarrow (x'_i, y'_i)$ from the reference and sensed images, the RMSE is computed according to:

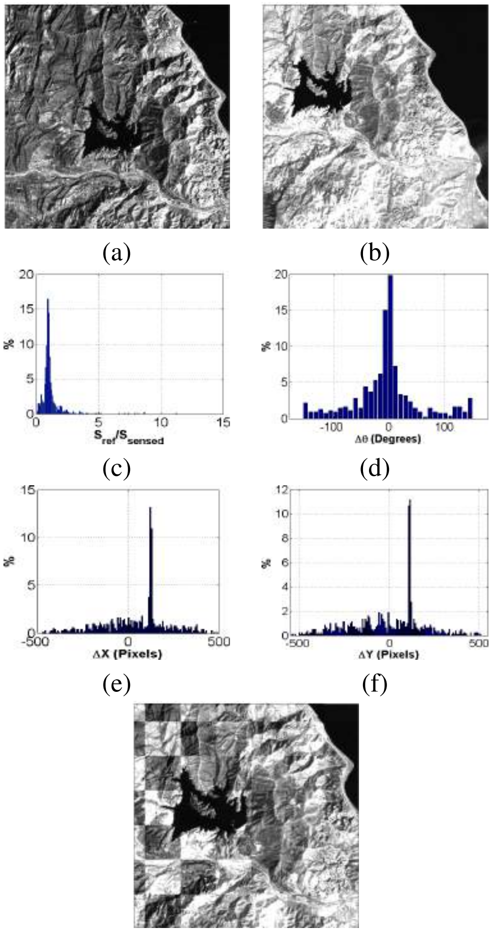
$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \tilde{x}_i)^2 + (y_i - \tilde{y}_i)^2}, \tag{5}$$

where $(\tilde{x}_i, \tilde{y}_i)$ denotes the transformed coordinates of (x'_i, y'_i).

We present the following results. Figs. 1(a) and 1(b) show two 600×600 Landsat images taken in 1984 (band 5) and 1986 (band 7), respectively. The underlying transformation clearly consists of substantial translation and negligible rotation and scale. The SIFT threshold gave rise to 797 and 833 key points for the reference and sensed images, respectively. Figs. 1(c) and 1(d) depict the scale ratio and rotation difference histograms (with bin widths of 0.075 and $9°$), respectively. The modes are easily located at $s_{\text{mode}} = 0.975$ and $\Delta\theta_{\text{mode}} = 1.33°$. (The exact mode location is determined via proper interpolation.) Figs. 1(e) and 1(f) show the histograms of the horizontal and vertical shifts, computed by (1) and (2), with equal bin widths of 7.5 pixels. The modes obtained at $\Delta x_{\text{mode}} = 124.62$ and $\Delta y_{\text{mode}} = 112.78$ are clearly evident.

Employing the outlier filters (3) and (4), with horizontal and vertical shift thresholds of a bin size results in 82 points (out of 797 correspondences), i.e., a rejection of $\sim 90\%$. Finally, employing the 1-step OLS yields the bottom-line transformation

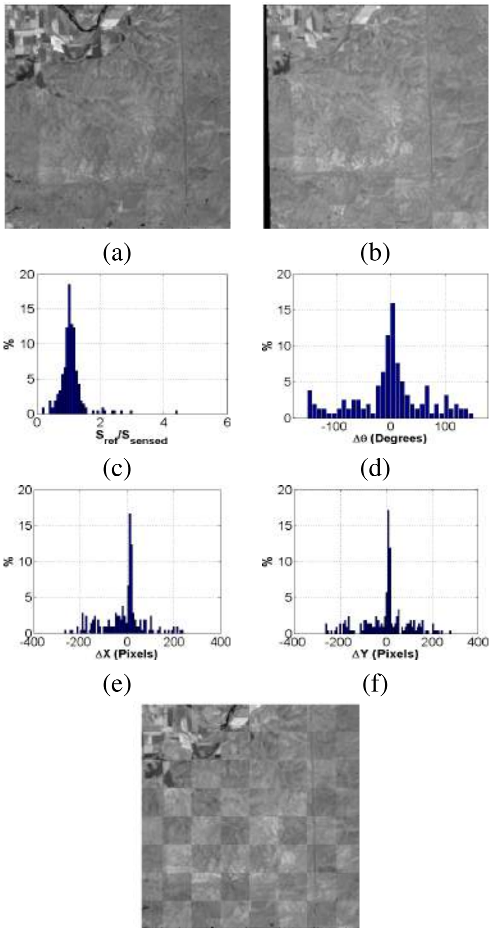
$< s, \theta, t_x, t_y >=< 0.99, 0.28°, 127.8, 112.34 >$. The RMSE in this case was 0.75 pixels (for 10 pairs picked manually). Fig. 1(g) illustrates the registration output (in the overlapping area) of the two images, which are superimposed on the same coordinate system. The above registration took 2.61[sec] on an old PC (Intel Q8200 with 3[GB] RAM and Vista OS). Note that the above histogram bin widths were fixed throughout all of our experiments.



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

Fig. 1. (a)-(b) Reference and sensed Landsat images (source: UCSB website of image processing and vision research labs), (c)-(f) histograms of scale ratio, orientation difference, and horizontal and vertical shifts, and (g) superimposed images after registration.

Figs. 2(a) and 2(b) depict an image pair of size 344×336 over the Konza site acquired by Landsat/ETM+ and IKONOS in the near infra-red band (NIR), respectively. (Wavelet decomposition was utilized here to bring the data to a similar spatial resolution, i.e., the IKONOS image was transformed to a spatial resolution of 32 meters, applying three levels of decomposition. Thus, the scaling expected for the given images is roughly 1.07.) In this case we had 209 initial SIFT correspondences. Figs. 2(c) and 2(d) depict the scale ratio and rotation difference histograms, respectively. Similarly, Figs. 2(e) and 2(f) show the corresponding histograms of the horizontal and vertical shifts, respectively. Fig. 2(g) depicts the registration result. The transformation obtained was $< s, \theta, t_x, t_y >=< 1.061, -0.00°, 13.56, 12.29 >$; this was computed from 45 inliers (21% of the initial number of correspondences). The RMSE in this case was 0.84 pixels and the run-time on our standard PC was 0.97[sec].



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

Fig. 2. (a)-(b) Reference and sensed images over Konza (source: MODIS Validation Core Sites), (c)-(f) histograms of scale ratio, orientation difference, and horizontal and vertical shifts, and (g) superimposed images after registration.

V. Failure Analysis

As indicated, we have tested our algorithm on a diverse set of multi-temporal, multi-spectral, and multi-sensor image pairs. We observed 76 successful registrations out of 94 trials (according to the 1-pixel RMSE criterion), i.e., a success rate of over 80%. Table I summarizes our results according to each image pair category (average RMSE and run-time are applicable only for successful registrations). Rows 2-4 demonstrate the algorithm's capability of handling, to a satisfactory extent, multi-band and multi-sensor image pairs.

As explained above, verifying the correctness of the transformation obtained was based upon the selection of GT points once the algorithm completes running. It would be of interest, of course, to devise an alternative, automatic measure, which will indicate whether the algorithm succeeds or fails in finding an appropriate transformation. We made use of the number of resulting inliers in our filter, for this purpose, in the following manner. Fig. 3 depicts the number of resulting inliers for each of the 94 registration trials; note the obvious gap between 4 to 7 inliers. The important observation is that all (of our) registration trials with less than 4 inliers resulted in a registration failure, while those with 6 inliers or more resulted in a successful registration (according to the RMSE criterion); thus we can exploit the number of resulting inliers for the desired success/failure indication. The above threshold of the number of inliers might need, of course, to be adjusted differently for other datasets. It represents a tradeoff for classifying a successful registration as a failure (higher threshold) and vice-versa (lower threshold). Registration failures were usually due to cardinal changes between the images (e.g., clouds covering a substantial area in one of the images) or reverse intensities in the images taken by different sensors at different spectral bands. We assessed the quality of SIFT key points in our failed registration trials by picking manually 8–10 GT point pairs and computing, in each case, the resulting transformation using a standard OLS procedure. Next, we applied this transformation to each SIFT key-point in the sensed image and computed the resulting Euclidean distance, d, to its corresponding nearest-neighbor SIFT key-point in the reference image. A correspondence is declared true if $d < 2$ [pix] and false, otherwise. We observed that in all failures there were at most 4 true correspondences (usually none), while in successful registrations there were at least 14 true correspondences. Theoretically, 2 true correspondences are sufficient to compute the parameters of the similarity transformation. However, since our algorithm is statistical in nature (i.e., uses quantized histograms), it needs a larger sample than this minimal number of correspondences which can be picked only in a deterministic, manual manner.

TABLE I
SUMMARY OF MS-SIFT REGISTRATION PERFORMANCE.

Category of Img. Pair	**# Succ.**	**# Fails.**	**Succ. [%]**	**<RMSE> [pix]**	**<Time> [sec]**
Multi-Temporal	53	10	84.12	0.718	0.849
Multi-Band	14	4	77.78	0.722	4.560
Multi-Sensor	6	0	100	0.838	0.950
Multi-Sensor & Multi-Band	3	4	42.86	0.743	0.840
Total	**76**	**18**	**80.8**	**0.729**	**1.24**

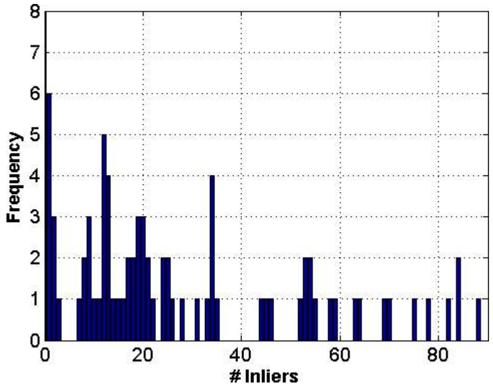


Fig. 3. Resulting inlier distribution.

VI. ENHANCEMENTS

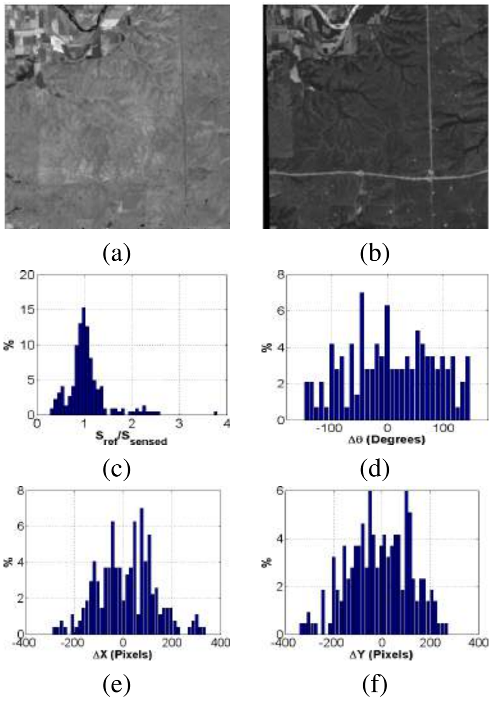
For failures caused by reverse intensities, various image processing enhancement techniques could be employed to improve the results. Consider the following data acquired over the Konza area to illustrate this. Figs. 4(a) and 4(b) show a (reference) NIR Landsat/ETM+ image and a (sensed) IR IKONOS image, respectively. Both images are of size 344×336. Due to the different spectral bands, dark areas in the reference image appear bright in the sensed image and vice-versa (see, for example the "+" shaped roads across the images and the river at the upper left corner of the images). Such intensity changes could make the SIFT correspondences unreliable, as correspondence computation relies on gradient orientations rather than mere gradient magnitudes. Figs. 4(c)-4(f) show the histograms of the SIFT characteristics obtained in this case. As opposed to the results in Fig. 2, these histograms exhibit no evident, single mode (except for the scale ratio histogram). We obtained here only a single inlier, which resulted, of course, in a registration failure.

To improve these results, we first sharpened both images according to the operator

$$\widetilde{f}(x,y) = f(x,y) - k\nabla^2 f(x,y), \tag{6}$$

where $\widetilde{f}(x,y) = f(x,y)$ is the sharpened image, $f(x,y)$ is the input image, $k \in (0,1]$, and $\nabla^2 f(x,y)$ is the discrete Laplacian operator. We then reversed the intensity, r, of the reference image (i.e., $T(r) = 1 - r$, assuming that $r \in [0,1]$). These two steps result in more reliable SIFT key-point correspondences. The sharpening enforces more dominant, true features at the detection stage of the difference of Gaussians (DoG), and a simple transformation is applied to overcome the problem of reverse intensities. Figs. 5(a) and 5(b) show the images after these enhancement steps. It is apparent that, visually, the images are much more alike. Figs. 5(c)-5(f) show the histograms of the various SIFT characteristics. In contrast to the histograms in Fig. 4, the modes of these histograms are unique and much more distinctive. Consequently, we obtained a successful registration due to 17 true correspondences (out of 154 initial ones). The RMSE value was 0.76 [pix]. Fig. 5(g) shows the registration result. We applied the above enhancements to 3 additional failed registrations with the same image characteristics (i.e., Landsat NIR vs. IKONOS IR) and got similar results. Of course, this method might not apply "as-is", and it is likely that a-priori knowledge (regarding, e.g., the acquiring sensors) would be required to tune its parameters, but it is reasonable to assume that such information should be available.

For comparison, we implemented also SIFT-based RANSAC variants, tested them on our dataset, and compared their performance to our method. Among the 94 image pairs tested, only 66 were registered successfully due to RANSAC with NN pruning (versus 76 due to MS-SIFT). In terms of accuracy, the average RMSE values were 0.97 [pix] and 0.69 [pix] due to RANSAC and MS-SIFT, respectively. This is mainly because RANSAC does not always distinguish correctly between inliers and outliers, which affects the accuracy of the resulting transformation. Finally, the average run-times per image pair were 1.44 [s] and 0.92 [s] due to RANSAC and MS-SIFT, respectively. This is due to the fact that RANSAC is more cumbersome and requires more computations, especially for a small fraction of inliers.



(a) (b)

(c) (d)

(e) (f)

Fig. 4. (a)-(b) Reference and sensed images over Konza (source: MODIS Validation Core Sites) and (c)-(f) histograms of scale ratio, orientation difference, and horizontal and vertical shifts.

VII. Conclusions

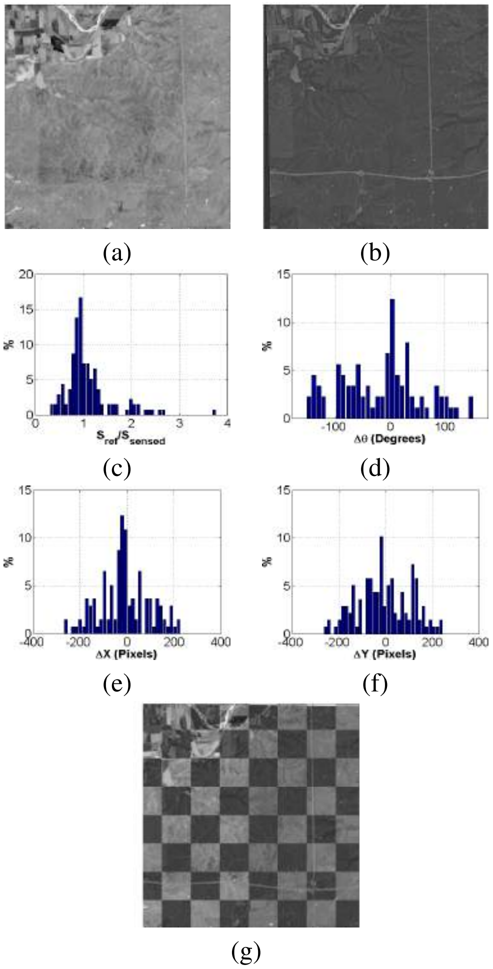
We presented here a simple SIFT-based variant for IR of remotely sensed images. Our method performs mode seeking in 4-D space (assuming a similarity transformation), followed by effective pruning of outlying SIFT key-point correspondences. Extended empirical studies on a diverse set of multi-temporal, multi-spectral, and multi-sensor images revealed good performance. The algorithm presented is fast and achieves subpixel accuracy. Also, we provided an automatic indicator as to the correctness of the registration obtained. Finally, we offered prospective enhancements for dealing with a failed registration. As part of future work, it would be of interest to explore the applicability of our MS-SIFT approach for more complex transformations and other descriptor types (e.g., SURF, GLOH, etc.).

References

[1] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *International J. of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] N. S. Netanyahu, J. Le Moigne, and J. G. Masek, “Georegistration of Landsat data via robust matching of multiresolution features,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 7, pp. 1586–1600, 2004.

[3] D. M. Mount, N. S. Netanyahu, and S. Ratanasanya, “New approaches to robust, point-based image registration,” in *Image Registration for Remote Sensing*, J. LeMoigne, N. S. Netanyahu, and R. D. Eastman, Eds. Cambridge University Press, March 2011.



(a) (b)

(c) (d)

(e) (f)

(g)

Fig. 5. (a)-(b) Images over Konza after enhancements, (c)-(f) histograms of their SIFT characteristics, and (g) registration result.

[4] Q. Li, G. Wang, J. Liu, and S. Chen, “Robust scale-invariant feature matching for remote sensing image registration,” *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 2, pp. 287–291, 2009.

[5] M. Teke, M. F. Vural, A. Temizel, and Y. Yardımcı, “High-resolution multispectral satellite image matching using scale invariant feature transform and speeded up robust features,” *Journal of Applied Remote Sensing*, vol. 5, no. 1, pp. 053 553–1–053 553–9, 2011.

[6] A. Sedaghat, M. Mokhtarzade, and H. Ebadi, “Uniform robust scale-invariant feature matching for optical remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 11, pp. 4516–4527, 2011.

[7] Q. Li, H. Zhang, and T. Wang, “Multispectral image matching using rotation-invariant distance,” *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 3, pp. 406–410, 2011.

[8] M. Fischler and R. Bolles, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography,” *Comm. of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[9] M. Hasan, X. Jia, A. Robles-Kelly, J. Zhou, and M. R. Pickering, “Multi-spectral remote sensing image registration via spatial relationship analysis on sift keypoints,” in *IEEE International Geoscience and Remote Sensing Symposium*, 2010, pp. 1011–1014.

[10] M. Hasan, M. R. Pickering, and X. Jia, “Modified SIFT for multi-modal remote sensing image registration,” in *IEEE International Geoscience and Remote Sensing Symposium*, 2012, pp. 2348–2351.

[11] B. Kupfer, “*A SIFT-Based Image Registration Algorithm for Remotely Sensed Data*,” M.Sc. Thesis, Bar-Ilan University, Israel, 2013, www.cs.biu.ac.il/∼nathan/registration/kupfer_thesis.pdf.