Abstract—In this letter, we propose a probabilistic method for the feature matching of remote sensing images which undergo nonrigid transformations. We start by creating a set of putative correspondences based on the feature similarity and then focus on removing outliers from the putative set and estimating the transformation as well. This is formulated as a maximum likelihood estimation of a Bayesian model with latent variables indicating whether matches in the putative set are inliers or outliers. We impose nonparametric global geometrical constraints on the correspondence using Tikhonov regularizers in a reproducing kernel Hilbert space. We also introduce a local geometrical constraint to preserve local structures among neighboring feature points. The problem is solved by using the expectation–maximization algorithm, and the closed-form solution of the transformation is derived in the maximization step. Moreover, a fast implementation based on sparse approximation is given which reduces the method computation complexity to linearithmic without performance sacrifice. Extensive experiments on real remote sensing images demonstrate accurate results of the proposed method which outperforms current state-of-the-art methods, particularly in case of severe outliers.

Index Terms—Feature matching, outlier, probabilistic inference, regularization, remote sensing.

I. INTRODUCTION

Feature matching, which refers to establishing reliable correspondence between two sets of features (particularly point features), is a critical prerequisite in feature-based registration, and it has a wide range of applications in remote sensing, including environment monitoring, change detection, image fusion, image mosaic, and map updating [1]–[3]. The features to be matched are typically extracted from two images of the same scene taken from different viewpoints, at different times, or by different sensors.

A popular strategy for solving the feature matching problem is to use a two-stage process [4]–[8]. In the first stage, a set of putative correspondences are computed by using a similarity constraint, which requires that points can only match points with similar local descriptors (e.g., scale invariant feature transform [SIFT] [9], [10]). This putative correspondence set typically includes most of the true matches, the inliers, and also a large number of false matches, or outliers, due to ambiguities in the similarity constraint. The second stage is designed to remove the outliers by using a geometrical constraint, which requires that the matches satisfy an underlying geometrical requirement. The inliers and the geometric parameters of the transformation are then obtained accordingly. Examples of this strategy include the classic RANdom SAmple Consensus (RANSAC) algorithm [11] and analogous robust hypothesize-and-verify methods [12]. They estimate a global transformation and reject outliers simultaneously. A correspondence is considered as an outlier and deleted if it is inconsistent with the transformation that the majority agree. Although these methods are very successful in many situations, they have had limited success if the geometrical constraints are nonparametric (e.g., nonrigid), and they also tend to degrade badly if the proportion of outliers in the putative correspondence set becomes large [13]. To address these problems, recently, some new nonparametric model-based methods have also been introduced, such as the graph shift [14] and identifying correspondence function (ICF) [13]. However, the matching accuracy of these methods is still not satisfying in case of large outlier percentage, particularly for matching remote sensing images.

It is a challenging task to generate accurate matches from remote sensing images. First, remote sensing images often contain local distortions caused by ground relief variations and imaging viewpoint changes, which means that they are not “exactly matchable” via a simple parametric model (e.g., rigid or affine transformation) as used in most existing methods. Therefore, high-dimensional nonrigid transformations are required to produce accurate alignments. Second, the complex nature of remote sensing images (e.g., unavoidable noise, occlusions, repeated structures, and multisensor data) often results in a high number of false matches, which have a significant impact on determining the transformational model. Therefore, a robust procedure of outlier removal is desirable. Third, for
large remote sensing images, the scale of extracted feature points is usually extremely large, e.g., tens of thousands. This poses a significant burden on typical feature matching methods, particularly in the nonrigid case. Therefore, it is of particular advantage to develop a more efficient technique.

In this letter, we propose a probabilistic method to address the aforementioned issues. The method is efficient and can handle nonrigid deformations in the image pairs, and it is also robust to a very large number of outliers. More precisely, we introduce a maximum likelihood framework for the robust estimation of transformation from a set of putative correspondences contaminated by noise and outliers. Our approach associates each correspondence with a latent variable which indicates whether it is an inlier or not, and then, it alternatively recovers the underlying transformation and estimates the inlier set by using an expectation–maximization (EM) algorithm. We model the transformation in a functional space, called the reproducing kernel Hilbert space (RKHS) [15]. The global smoothness constraint is imposed on the transformation by using Tikhonov regularizers, and we also introduce a local geometrical constraint to preserve local structures among neighboring feature points. Moreover, we develop a fast implementation based on a similar kind of idea as the subset of regressors method [16].

II. ALGORITHM

This section describes the proposed feature matching algorithm for remote sensing images. We start by laying out the maximum likelihood formulation and then introduce the global and local geometrical constraints for feature matching. We subsequently give our global–local feature matching algorithm and provide a fast implementation. Finally, we analyze the computational complexity of the proposed approach.

A. Maximum Likelihood Formulation

Given a set of putative correspondences \( S = \{ (x_n, y_n) \}_{n=1}^N \) which may contain some unknown outliers, where \( x_n \) and \( y_n \) are the 2-D spatial positions of two corresponding points, our goal is to distinguish inliers from the outliers by robustly fitting a transformation (e.g., the displacement function) \( f \) for the underlying inliers, i.e., for an inlier correspondence \((x_n, y_n)\), \( y_n = x_n + f(x_n) \). For generality, \( f \) is supposed to be nonrigid and is modeled by requiring it to lie within a specific functional space \( \mathcal{H} \), namely, an RKHS [15]. The field \( f \) then takes an explicit kernel representation

\[
f(x) = \sum_{n=1}^{N} \Gamma(x, x_n)c_n
\]

where \( \Gamma \) is a positive definite matrix-valued kernel defining the RKHS \( \mathcal{H} \), e.g., \( \Gamma(x, x) = \kappa(x, x) \cdot I = e^{-\beta \|x-x\|^2} \cdot I \) with \( \beta \) determining the width of the range of interaction between samples (i.e., neighborhood size), and \( \{c_n\}_{n=1}^N \) is the set of unknown coefficients.

To have a robust estimation, we build a model that includes the outliers [4], [3], [17]. It assumes the noise of the inlier to be Gaussian with zero mean and uniform standard deviation \( \sigma \), and the outlier is assumed to be uniform \( 1/\alpha \) with \( \alpha \) being the area of the input image. We then associate the \( n \)th correspondence with a latent variable \( z_n \in \{0,1\} \), where \( z_n = 1 \) indicates that the correspondence \((x_n, y_n)\) is an inlier and \( z_n = 0 \) points to an outlier. Thus, the likelihood is a mixture model given by

\[
p(Y|X, \theta) = \prod_{n=1}^{N} p(y_n, z_n|x_n, \theta)
\]

\[
= \prod_{n=1}^{N} \left( \frac{\gamma}{2\pi\sigma^2} e^{-\frac{\|y_n - x_n - f(x_n)\|^2}{2\sigma^2}} + \frac{1 - \gamma}{\alpha} \right) \tag{2}
\]

where \( X = (x_1, \ldots, x_N)^T \) and \( Y = (y_1, \ldots, y_N)^T \) are the feature points, \( \theta = \{f, \sigma^2, \gamma\} \) includes a set of unknown parameters, and \( \gamma \) is the mixing coefficient.

We give a maximum likelihood estimation of the parameter set \( \theta \), i.e., \( \theta^\text{old} = \arg \max \theta p(Y|X, \theta) \), which is equivalent to minimizing its negative log-likelihood function. This can be solved by using an EM approach, where the complete-data log posterior is given by

\[
Q(\theta, \theta^\text{old}) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} p_n \|y_n - x_n - f(x_n)\|^2 - \ln \sigma^2 \sum_{n=1}^{N} p_n
\]

\[
+ \ln \gamma \sum_{n=1}^{N} p_n + \ln(1 - \gamma) \sum_{n=1}^{N} (1 - p_n) \tag{3}
\]

where \( p_n = P(z_n = 1|x_n, y_n, \theta^\text{old}) \) is a posterior probability, which indicates to what degree \((x_n, y_n)\) is an inlier.

E-Step: Denote \( P = \text{diag}(p_1, \ldots, p_N) \) as a diagonal matrix, where \( p_n \) can be computed by using the current parameter set \( \theta^\text{old} \) based on Bayes’ rule

\[
p_n = \frac{\gamma e^{-\|y_n - x_n - f(x_n)\|^2/2\sigma^2} + 1/\alpha}{\gamma e^{-\|y_n - x_n - f(x_n)\|^2/2\sigma^2} + (1 - \gamma) \sum_{n=1}^{N} 1/\alpha} \tag{4}
\]

M-Step: Re-estimate the parameters using current responsibilities: \( \theta^\text{new} = \arg \max \theta Q(\theta, \theta^\text{old}) \). Taking derivatives of \( Q \) with respect to \( \sigma^2 \) and \( \gamma \) and setting them to zero, we obtain

\[
\sigma^2 = \frac{\text{tr} \left( (Y - X - F)^T P (Y - X - F) \right)}{2 \cdot \text{tr} (P)} \tag{5}
\]

\[
\gamma = \frac{\text{tr} (P)}{N} \tag{6}
\]

where \( F = (f(x_1), \ldots, f(x_N))^T \). In order to complete the EM algorithm, the transformation \( f \) should be estimated in the M-step. We will discuss it later in the following sections.

After the EM iteration converges, with a predefined threshold \( \tau \), the inlier set \( I \) is obtained according to the criterion

\[
I = \{ (x_n, y_n) : p_n > \tau, n \in \mathbb{N} \}. \tag{7}
\]

B. Global and Local Regularizations

Considering the terms in (3) related to the transformation \( f \), we have \( Q(f) = -(1/2\sigma^2) \sum_{n=1}^{N} p_n \|y_n - x_n - f(x_n)\|^2 / 2 \). This is, in general, ill-posed since the \( f \) is not unique. To have a stable solution, a typical strategy is to impose regularization on \( f \) to control the complexity of the hypothesis space, such as the Vapnik–Chervonenkis dimension [18]. Here, we consider the Tikhonov regularization [19], and the regularization term is given by

\[
\phi(f) = \|f\|^2 = C^T \Gamma C \tag{8}
\]
where \( C = (c_1, \ldots, c_N)^T \) is the coefficient set of \( f \) and the kernel matrix \( \Gamma \in \mathbb{R}^{N \times N} \) is the so-called Gram matrix with \( \Gamma_{ij} = \kappa(x_i, x_j) = e^{-\beta|x_i - x_j|^2} \).

The Tikhonov regularization is a global constraint, and it is useful to keep the overall spatial connectivity of the point correspondences during matching. However, for feature matching, the local structures among neighboring feature points are also very strong and stable. This is particularly beneficial when the motion between the two images is nonrigid. In this case, a local geometrical constraint is desired to establish accurate correspondences. We preserve the local neighborhood structure during matching.

In our problem, we hope that the local structures in \( Y \) could be preserved after the displacement of \( X \), which could be achieved by the following three steps [3]. First, search the \( K \) nearest neighbors for each point in \( X \), and enforce the weight \( W_{ij} = 0 \) if \( x_j \) does not belong to the set of neighbors of \( x_i \), where \( W \) is an \( N \times N \) weight matrix with \( W_{ij} \) summarizing the contribution of \( x_j \) to \( x_i \) for reconstruction. Second, minimize the reconstruction errors measured by the cost function

\[
E(W) = \sum_{i=1}^{N} \left\| x_i - \sum_{j=1}^{N} W_{ij} x_j \right\|^2 \tag{9}
\]

under a constraint that the rows of the weight matrix sum to one: \( \sum_{j=1}^{N} W_{ij} = 1 \). The optimal weights \( W_{ij} \) can be obtained by solving a least squares problem. Third, the local geometry of each inlier point after the transformation \( f \) is preserved by minimizing the cost function

\[
E(C) = \sum_{i=1}^{N} p_i \left( x_i^T + \Gamma_i, C \right) - \sum_{j=1}^{N} W_{ij} \left( x_j^T + \Gamma_j, C \right) \tag{10}
\]

where \( \Gamma_i \) is the \( i \)th row of \( \Gamma \), and \( p_i \) is a posterior probability which plays a role of soft assignment to determine the inliers.

C. Transformation Estimation

Combining the global regularization term in (8) and the local regularization term in (10), the coefficient set \( C \) of the transformation \( f \) can be solved by minimizing

\[
Q(C) = \frac{1}{2\sigma^2} \sum_{n=1}^{N} p_n \| y_n - x_n - \Gamma_n, C \|^2 + \frac{\lambda}{2} C^T \Gamma C
+ \frac{\eta}{2} \sum_{i=1}^{N} p_i \left( x_i^T + \Gamma_i, C \right) - \sum_{j=1}^{N} W_{ij} \left( x_j^T + \Gamma_j, C \right) \tag{11}
\]

where the positive real numbers \( \lambda \) and \( \eta \) control the tradeoff between data fitting and global/local geometric preserving.

To solve \( C \), we take the derivative of (11) with respect to \( C \) and set it to zero; then, we obtain a linear system

\[
(P\mathbf{T} + \lambda \sigma^2 \mathbf{I} + \eta \sigma^2 Q \mathbf{I}) C = P\mathbf{Y} - (\mathbf{P} + \eta \sigma^2 Q) X \tag{12}
\]

where \( Q = (I - W)\mathbf{T}(I - W) \). Until now, all the parameters have been solved. We summarize our approach in Algorithm 1.

Algorithm 1: The Proposed Algorithm

**Input:** Correspondence set \( S = \{ (x_n, y_n) : n \in \mathbb{N}_N \} \), kernel \( \Gamma \), parameters \( \lambda, \eta, \tau \)

**Output:** Inlier set \( \mathcal{I} \)

1. Initialize \( \gamma \), \( \mathbf{C} = 0 \), \( \mathbf{P} = \mathbf{I} \times \mathbf{X} \times \mathbf{N} \);
2. Set \( a \) to the area of the image;
3. Initialize \( \sigma^2 \) by Eq. (5);
4. Construct the Gram matrix \( \Gamma \) using the definition of \( \Gamma \);
5. Search the \( K \) nearest neighbors for each point in \( X \);
6. Compute \( W \) by minimizing the cost function (9);
7. \textbf{repeat}
   8. \textbf{E-step:} \( \mathbf{P} = \text{diag}(p_1, \ldots, p_N) \) by Eq. (4);
   9. \textbf{M-step:}
      10. Update \( \mathbf{C} \) by solving linear system (12);
      11. Update \( \sigma^2 \) and \( \gamma \) by Eqs. (5) and (6);
    \textbf{until} \( Q \) converges;
8. The consensus set \( \mathcal{I} \) is determined by Eq. (7).

D. Fast Implementation

In the feature matching problem, the point set typically contains hundreds or thousands of points, which causes significant complexity problems (in time and space). Consequently, we adopt a sparse approximation and randomly pick only a subset of size \( M \) input points \( \{ x_m \}_{m=1}^{M} \) to have nonzero coefficients in the expansion of the solution [i.e., (1)]. It has been experimentally proven that this approximation works well, and simply selecting a random subset of the input points in this manner performs no worse than more sophisticated and time-consuming methods [16]. Thus, we seek a solution of form

\[
f(x) = \sum_{m=1}^{M} \Gamma(x, x_m) c_m. \tag{13}
\]

The chosen point set \( \{ x_m \}_{m=1}^{M} \) is somewhat analogous to “control points.” By using the sparse approximation, the linear system (12) becomes

\[
(U^T P U + \lambda \sigma^2 \mathbf{I} + \eta \sigma^2 U^T Q U) C^s = U^T P Y - U^T (P + \eta \sigma^2 Q) X \tag{14}
\]

where the coefficient matrix \( C^s = (c_1, \ldots, c_M)^T \in \mathbb{R}^{M \times 2} \), the kernel matrix \( \Gamma^s \in \mathbb{R}^{M \times M} \) with \( \Gamma^s_{ij} = \kappa(x_i, x_j) = e^{-\beta||x_i - x_j||^2} \), and \( U \in \mathbb{R}^{N \times M} \) with \( U_{ij} = \kappa(x_i, x_j) = e^{-\beta||x_i - x_j||^2} \).

E. Computational Complexity

To search the \( K \) nearest neighbors for each point in \( X \), the time complexity should be close to \( O((K + N) \log N) \) by using the k-d tree [20]. According to (9), the time complexity of obtaining the weight matrix \( W \) is \( O(K^3 N) \) because each row of \( W \) can be solved separately with \( O(K^3) \) time complexity. Moreover, due to the Gram matrix being of size \( N \times N \), the time complexity of solving the linear system (12) is \( O(N^3) \). Since \( K \ll N \), the total complexity of our method can be written as \( O(N^3) \). The space complexity of our method scales like \( O(N^2) \) due to the memory requirements for storing the Gram matrix \( \Gamma \).
By using the sparse approximation, the size of the Gram matrix reduces to $M \times M$, and the time complexity of solving the linear system (14) reduces to $O(M^3 + M^2N + KMN + K^2N)$. Therefore, the total time complexity is $O(K^2N + M^2N + N \log N)$, which is about linearithmic with respect to the scale of the given correspondence set. The space complexity reduces to $O(MN + KN)$ due to the memory requirements for storing $U$ and $W$. This is significant for large-scale problems.

## III. EXPERIMENT RESULTS

We test the performance of our proposed algorithm and verify the global and local regularizations on real images. We compare our method with other three state-of-the-art feature matching methods such as RANSAC [11], ICF [13], and GS [14]. The parameters of the four methods are all fixed, and the experiments are performed on a laptop with 2.5-GHz Intel Core CPU, 8-GB memory, and MATLAB code.

### A. Data Sets and Settings

The test data consist of three types of image pairs with sizes from 638 × 750 to 1432 × 1632, which were captured over Tokyo, Japan, Wuhan, China, and Nantong, China. The first type contains 23 panchromatic aerial photographs of urban areas captured by an onboard camera of a helicopter in different height and direction. The second type contains 23 aerial photographs of mountain areas captured by an onboard camera of an unmanned aerial vehicle (UAV). The third contains 36 image pairs coming from different sources, where the reference images are the satellite-based synthetic aperture radar (SAR) data of RADARSAT II and the sensed images are the airborne SAR data of UAV. The image pairs involve ground relief variations and imaging viewpoint changes and hence are not exactly matchable via a parametric model such as rigid transformation.

We use the open source VLFeat toolbox [21] to determine the initial matches of SIFT [9], and all the matching methods then aim to remove the outliers contained in the initial SIFT matches. To establish the ground truth, i.e., determine the correctness of each correspondence, we use a method combining subjectivity and objectivity. We first use a matching algorithm to establish rough correspondences and then confirm the results artificially. The performance is characterized by precision and recall, where the precision is defined as the ratio of the preserved inlier number and the preserved correspondence number and the recall is defined as the ratio of the preserved inlier number and the inlier number contained in the putative correspondences.

**Parameter Setting:** There are mainly seven parameters in our method: $K$, $\lambda$, $\eta$, $\tau$, $\gamma$, $\beta$, and $M$. Parameter $K$ controls the number of nearest neighbors for linear reconstruction. Parameters $\lambda$ and $\eta$ control the tradeoff between data fitting and global/local geometric preserving. Parameter $\tau$ is a threshold, which is used for deciding the correctness of a correspondence. Parameter $\gamma$ reflects our initial assumption on the amount of inliers in the putative correspondence sets. Parameter $\beta$ determines how wide the range of interaction is between feature points, and parameter $M$ is the required number of control points for sparse approximation. We set $K = 15$, $\eta = 3$, $\lambda = 1000$, $\tau = 0.5$, $\gamma = 0.9$, $\beta = 0.1$, and $M = 15$.

### B. Results on Remote Sensing Images

In this section, we give quantitative comparisons of our method with RANSAC [11], ICF [13], and GS [14] on the 82 image pairs described in the last section.

The statistical results of the four methods are summarized in Fig. 1. The cumulative distribution of the initial inlier ratio is shown on the left of Fig. 1, where the average inlier percentage is about 41.09%. In the right figure, each scattered dot represents a precision–recall pair. For RANSAC, the fundamental matrix is chosen as the parametric model. It can produce a satisfying result when the initial correct match percentage in the putative set is high. ICF and GS usually have high precisions or recalls, but not simultaneously. In contrast, our proposed method has the best precisions and recalls which are concentrated on the top right corner.

We also test the average run times on the 82 remote sensing image pairs and report the results in Table I. We see that our method is the fastest one, followed by ICF. RANSAC is not very efficient due to the low initial inlier percentages in the data sets. GS is quite efficient when the scale of the data set is not large but degrades rapidly as the scale of the data set grows. The large variances of the initial inlier percentages lead to large variances of the run times for RANSAC and GS.

**TABLE I**

<table>
<thead>
<tr>
<th>Time (s)</th>
<th>12.37 ± 14.07</th>
<th>2.25 ± 2.42</th>
<th>12.95 ± 16.53</th>
<th>1.21 ± 1.18</th>
</tr>
</thead>
</table>

### C. Robustness Test

Next, we use a typical image pair to test the robustness of our method, as shown in the first row of Fig. 2. We also visualize the distribution of initial SIFT matches in the images. In our evaluation, we consider the following three scenarios: vary the inlier number, total number, and inlier ratio by adding random matches. First, the initial inlier percentage is clearly an important factor that influences the matching performance. Thus, we fix the inlier number as 100 and vary the inlier ratio from 0.01 to 0.3 at an interval of 0.01. Second, the absolute number of correct matches can also influence the performance. We fix the inlier ratio as 0.2 and vary the inlier number from 5 to 150 at an interval of 5 in the image pair. Third, in order to test the performance in the situation where both the inlier number and ratio are low, we fix the total number of the putative set as 500 and vary the inlier ratio from 0.01 to 0.3 at an interval of 0.01.
IV. Conclusion

Within this letter, we proposed a new probabilistic method for the feature matching of remote sensing images under non-rigid transformations. It simultaneously estimates the transformation and generates accurate correspondences by an iterative EM algorithm under a maximum likelihood framework with both global and local regularizations. The method is efficient and robust and is able to handle severe outliers. The qualitative and quantitative comparisons on several types of remote sensing image pairs demonstrate that our method outperforms other state-of-the-art methods, particularly in case of severe outliers.

REFERENCES