This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2961480, IEEE Transactions on Image Processing

A Context-aware Locality Measure for Inlier Pool Enrichment in Stepwise Image Registration

Su Zhang, Wanjing Zhao, Xuying Hao, Yang Yang, *Member, IEEE*, and Cuntai Guan, *Fellow, IEEE*

***Abstract*—We present a feature-based image registration method, the stepwise image registration (SIR), with a closed-form solution. Our SIR creates an inlier pool and a candidate pool as the initialization, and then gradually enriches the inlier pool and refines the transformation. In each step, the enriched correspondence exclusively tunes the transformation coefficient within the confirmed inlier pairs, instead of updating the mapping using the complete putative set. In turn, the refined transformation prunes inconsistent mismatches to alleviate the incoming matching ambiguity. The context-aware locality measure (CALM) is designed for dissimilarity measure. The capability of the CALM can be enhanced by the progressive inlier pool enrichment. Finally, a retrieval process is performed based on the finest CALM and alignment, by which the inlier pool is maximized. Extensive experiments of enrichment evaluation, feature matching, image registration, and image retrieval demonstrate the favorable performance of our SIR against state-of-the-art methods. The code and datasets are available at https://github.com/sucv/SIR.**

***Index Terms*—Feature matching, registration, dissimilarity measure, non-rigid.**

I. Introduction

FEATURE-based image registration [1] is the process of geometrically aligning the sensing image onto the reference image by recovering the correspondence of the image feature sets, where each feature is usually referred to as a discrete point. It has been widely used in many fields, such as computer vision, remote sensing, medical image analysis, image retrieval, and image mosaic.

The performance of this framework includes (i) putative correspondence establishment, (ii) mismatch removal, and (iii) image transformation, which can be affected by two issues. The first issue is the irreconcilable conflict between the sufficiency and correctness of the initially established putative correspondence. The nearest neighbor and distance ratio (NNDR) algorithm working on the feature descriptors, e.g., the scale-invariant feature transform (SIFT) descriptors [2], requires that a match is accepted only if the ratio of the nearest and second-nearest descriptor distances satisfies a threshold τ [3]. On the one hand, a loose threshold allows more features to be identified as putative inliers. Therefore, the subsequent processes are possible to preserve sufficient true inliers, and further model a fine image transformation. On the other hand, a strict threshold guarantees that the outlier ratio of the putative inliers will not exceed the tolerance of the mismatch removal methods, and also renders computational burden reasonable. Methods chosen to maintain sufficiency will then face the second issue, namely the matching ambiguity caused by outlier contamination and non-rigid distortion, which becomes worse when images suffer from low overlap ratio or large viewpoint changes. The ambiguous correspondences and unnecessary computation consumed by egregious outliers may deteriorate the mismatch removal process and make the complexity intractable.

To address the two issues mentioned above, we present a robust and efficient image registration method by exploiting the reciprocity between distinctive and ambiguous features. The following observations are the basis of our method. On the one hand, the alignment of inlier pairs with high certainty should reduce the measure ambiguity geometrically for other pairs. On the other hand, the inherent consistency enables the inliers to play as structure descriptors. Such descriptors will have more distinctiveness if they can locally surround a candidate, instead of locating far apart to the candidate. Accordingly, our method seeks to establish the alignment using the reliable inliers, and gradually enrich the inlier set to improve the dissimilarity measure.

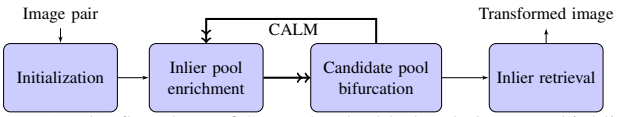

Figure 1: The flowchart of SIR. The double-headed arrows highlight the stepwise process. The main idea of our SIR is to interpret the image registration as a stair-climbing fashion, where the ground is the putative inliers, and the yield from the n-th stair can strengthen the capability of the CALM estimator, resulting in a more accurate measure for the next attempt.

More specifically, our method alternatively enriches the inlier pool and bifurcates the candidate pool in a stepwise manner, as shown in Fig. 1. During the initialization, a universal set for guaranteeing the sufficiency, and its subset with reliable correspondences, are obtained to build the candidate pool and the inlier pool, respectively. A robust dissimilarity measure named context-aware locality measure (CALM) is used to drive the stepwise process. It estimates a regularized mixture of neighborhood relationship, inter-neighborhood distance, and context information. The inlier pool is gradually enriched by candidates that are preserved by the CALM. Therefore its availability for neighborhood construction is enhanced owing

This work was supported by the National Nature Science Foundation of China [41971392]. (*Corresponding authors: Cuntai Guan; Yang Yang.*)

Su Zhang and Cuntai Guan are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore, 639798, Singapore (email: sorazcn@gmail.com; ctguan@ntu.edu.sg).

Wanjing Zhao, Xuying Hao, and Yang Yang are with the School of Information Science and Technology, Yunnan Normal University, Kunming, 650092, China (email: z_wanjing@126.com; hx_yying@126.com; yyang_ynu@163.com).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2961480, IEEE Transactions on Image Processing

to the more reasonable and abundant spatial distribution. A transformation modeled using the known inliers is applied to the candidate pool. Therefore the latter is bifurcated with reduced matching ambiguity. Finally, a retrieval process is performed to maximize the inlier pool based on the finest CALM and alignment.

The contributions of our work are twofold. First, a mathematical model is introduced to parameterize the stepwise process. The closed-form solution is obtained with reduced ambiguity. Second, a robust dissimilarity measure CALM is designed to remove mismatch based on the regularized mixture of neighborhood relationship, inter-neighborhood distance, and context information. With its capability being further enhanced by the inlier pool enrichment, the CALM ensures the desired functionality of the stepwise process. Extensive experiments on enrichment evaluation, feature matching, image registration, and image retrieval manifest that our method can yield better results comparing to state-of-the-art methods with the tractable time cost. Since the enrichment process goes like walking up stairs intuitively, we name the proposed method as stepwise image registration (SIR).

II. Related Works

Methods for feature-based image registration can be roughly categorized as traditional (non-deep learning) and deep learning based. In this section, we briefly review these existing methods that are relevant to our SIR.

The traditional methods using hand-crafted features usually recast the image registration as a re-sampling, graph matching, non-parametric interpolation, or point set registration problem. The random sample consensus (RANSAC) [4] and its variants, such as the maximum likelihood estimation sample consensus (MLESAC) [5], progressive sample consensus (PROSAC) [6] and randomized RANSAC [7], [8], are the representatives for the re-sampling approaches. These methods aim to obtain the smallest possible outlier-free subset by trials of re-sampling, they are vulnerable to outliers locating within the decision margin, or the non-rigidity violating the employed parametric model, e.g., the affine or homogeneous model. In the field of graph matching methods, representative studies such as spectral matching [9], dual decomposition [10], graph shift (GS) [11], and deformable graph matching (DGM) [12] have been presented. Graph matching methods mainly suffer the NP-hardness despite the robust matching and recognition performance. Recently, locality preservation matching (LPM) [13], [14] and its variant, the guided locality preservation matching (GLPM) [15] are presented. They have achieved promising results on mismatch removal with linearithmic time and linear space complexities. Non-parametric interpolation methods, such as identifying correspondence function (ICF) [16], bounded distortion (BD) [17] and vector field consensus (VFC) [18], commonly model a slow-and-smooth motion field [19] as the prior condition for interpolation. Point set registration methods remove mismatches by first aligning the feature point sets, and then obtaining the true matches according to the affinity matrix or distance threshold. Representative methods include iterate closest point (ICP) [20], thin-plate spline robust point matching (TPSRPM) [21], coherent point drift (CPD) [22], and global-local mixture distance (GLMD) [23], etc. Both of the non-parametric interpolation and point set registration methods are featured by the cubic complexities. Technically, our work belongs to the category of point set registration method.

The great success shown by deep learning [24] in various computer vision fields also motivates us to model the registration problem using deep learning framework. Great efforts have been made on 2-D image [25], 3-D Volumetric image [26] matching, and transformation estimation [27], [28], [29]. Other than feeding the regular 2-D pixel or 3-D voxel grids data to the deep net architecture, methods working directly on point sets with irregular format are recently proposed, which include but not limited to PointNet [30], PointNet++ [31], pointwise CNN [32] and ShapeContextNet [33]. Though the deep learning-based methods can yield promising results, they are demanding on the computational power, or usually required to be trained on a massive amount of data. Their interpretability [34] is still a topic of active research.

It is worth noting that the stepwise strategy typically utilized in traditional registration methods have multiple interpretations conceptually. The refinement interpretation is embedded by a coarse-to-fine insight. It starts with the whole point sets and iteratively rejects outliers in the alternating correspondence estimation and transformation updating, among which ICP [20] and CPD [22] are two of the most famous works. The uncertainty interpretation emerges from a probabilistic/statistics point of view. It measures the confidence of a solution given a particular set of parameters and conditions, and accordingly preserves the candidates that best consist with the applied constraints [35], [36], [37], [38], [39]. The fusion interpretation enriches the initial putative correspondences by combining different feature descriptors (e.g., the SIFT [2], speeded up robust features (SURF) [40], local intensity order pattern (LIOB) [41], etc.) that capture diverse visual evidences, some of the outstanding works are reported by Hu et. al [42], [43]. The propagation interpretation, to which our SIR belongs, initializes a small reliable set and then gradually grows the seeds by adding the remaining candidates under certain criteria. PROSAC [6] and sequential correspondence verification (SCV) [44], [45] are some of the outstanding works within this context.

Technically, the GLPM [15] is the published work most relevant to our SIR. The GLPM establishes the neighborhood in the view of graph matching [46] using a small and reliable set, and then removes outliers by comparing the neighborhood consistency of each pair from the universal set in one-shot. As a descriptor, however, the established neighborhood featuring limited and fixed distribution may not be sufficiently distinctive. This issue can be further escalated by the binary distance used for dissimilarity measure since it overlooks the spatial consistency. Our SIR can be taken as a gradually generalized version of the GLPM with an improved continuous dissimilarity measure.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2961480, IEEE Transactions on Image Processing

III. Stepwise Image Registration (SIR)

Our SIR aims to realize image registration in a stepwise manner (see notation[1]). In this section, we first formulate the problem and derive the closed-form solution. We then elaborate the context-aware locality measure (CALM) and candidate bifurcation and finally provide the main process and pseudo-code.

A. Problem Formulation

Given I^s and I^r the sensed and reference images, the universal set $\boldsymbol{\mathcal{S}} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ comprising the source point sets $\mathbf{X} = \{\mathbf{x}_i | i = 1, 2, ..., N\}$ and the target point sets $\mathbf{Y} = \{\mathbf{y}_i | i = 1, 2, ..., N\}$ is obtained using the SIFT extractor and NNDR. Our goal is to obtain the transformed image I^o by recovering the geometrical transformation $\mathbf{f} : I^r = \mathbf{f}(I^s, \boldsymbol{\theta}^*)$ registering I^s onto I^r, where $\boldsymbol{\theta}^*$ is the unknown transformation coefficient.

A straightforward idea is to model $\mathbf{f}$ based on an inlier set $\boldsymbol{\mathcal{I}}^* = \arg\min_{\boldsymbol{\mathcal{I}}} Q(\boldsymbol{\mathcal{I}}; \boldsymbol{\mathcal{S}}, \lambda)$, and the objective function is therefore formulated as:

$$Q(\boldsymbol{\mathcal{I}}; \boldsymbol{\mathcal{S}}, \lambda) = \sum_{i|\mathbf{x}_i, \mathbf{y}_i \in \boldsymbol{\mathcal{S}}} C^{\boldsymbol{\mathcal{S}}}(\mathbf{x}_i, \mathbf{y}_i) - \lambda |\boldsymbol{\mathcal{I}}|, \tag{1}$$

where $C^{\boldsymbol{\mathcal{S}}}$ is the global dissimilarity measure, penalizing any pairs which do not preserve the point-to-point distance within their own point set. The second term maximizes the inliers with λ controlling the strength and $|\cdot|$ denoting the cardinality of a set. Minimizing Q is equivalent to maximizing the number of inlier pairs and minimizing their dissimilarity measure C.

To improve the correctness of the dissimilarity measure, Q is rewritten based on the following considerations. First, the rigidity (planar translation and rotation) favors the reliability of $C^{\boldsymbol{\mathcal{S}}}$, under which the consistency like the absolute distances of points are preserved. However, if non-rigid deformation is present, such preservation will not hold, especially for points that are far apart [46]. A measure $C^{\boldsymbol{\mathcal{N}}}$ concerning only the neighborhood structure is feasible since the local consistency is reliable [46], [47]. Second, if the major components of the neighborhood are outliers, an inlier pair might be falsely rejected owing to the erroneously high cost. Hence the measure $C^{\boldsymbol{\mathcal{N}} \subset \boldsymbol{\mathcal{I}}}$ should require that the neighborhood structure is built only by neighboring inliers. Third, the distinctiveness of the neighboring inliers may be problematic when the candidates to be measured are distant. In this case, the measure is easier to produce false positive as many distant candidates have similar neighboring inliers. To address this issue, a stepwise feature matching strategy should be adopted for progressive enrichment of the inlier pool instead of the monolithic matching based on a fixed yet limited inlier set. Hopefully, the sequentially recovered inliers will pose evenly and thereby improve the accuracy. Meanwhile, a continuous local dissimilarity measure is also desired instead of the binary one [48], [13], [14], [15]. Fourth, a transformation $\mathbf{f}$ can benefit the matching process if it is modeled by the inlier pairs. As the stepwise feature matching goes, the structural discrepancies of inlier and outlier pairs from the candidate pool will be bifurcated, resulting in the reduction of matching ambiguity. Suppose that the stepwise matching process features M iterations, each of which specifically finds a sub-inlier set $\boldsymbol{\mathcal{I}}_m$, we can therefore rewrite Q in the form:

$$Q(\boldsymbol{\mathcal{I}}; \boldsymbol{\mathcal{S}}, \lambda) = \sum_{m=1}^{M} \left(\sum_{i|\hat{\mathbf{x}}_i, \mathbf{y}_i \in \boldsymbol{\mathcal{C}}_m} C^{\boldsymbol{\mathcal{N}} \subset \boldsymbol{\mathcal{I}}}(\hat{\mathbf{x}}_i, \mathbf{y}_i) - \lambda |\boldsymbol{\mathcal{I}}_m| \right), \tag{2}$$

where $C^{\boldsymbol{\mathcal{N}} \subset \boldsymbol{\mathcal{I}}}$ implies that the dissimilarity measure is based on the neighboring inliers, e.g., the K nearest inliers, and $\hat{\mathbf{x}}_i = \mathbf{f}(\mathbf{x}_i)$ is the transformed feature point. It should be noted that both the neighboring inliers and the transformation $\mathbf{f}$ is constructed based on the inlier pool obtained at the $(m-1)$-th iteration, and when $m = 1$, they are acquired using the initial inlier pool. Therefore they will not affect the optimization of Q as unknown variables explicitly.

To find the closed-form solution of Q, let us introduce a binary vector $\mathbf{p} = \{p_i \in 0, 1 | i = 1, 2, ..., N\}$ with each entry indicating the matching correctness of the i-th point pair, and then substitute $|\boldsymbol{\mathcal{I}}_m|$ by $\sum_{\hat{\mathbf{x}}_i, \mathbf{y}_i \in \boldsymbol{\mathcal{C}}_m} p_i$ into Eq. 2 and obtain:

$$Q(\mathbf{p}; \boldsymbol{\mathcal{S}}, \lambda) = \sum_{m=1}^{M} \left(\sum_{i|\hat{\mathbf{x}}_i, \mathbf{y}_i \in \boldsymbol{\mathcal{C}}_m} p_i \left(C^{\boldsymbol{\mathcal{N}} \subset \boldsymbol{\mathcal{I}}}(\hat{\mathbf{x}}_i, \mathbf{y}_i) - \lambda \right) \right). \tag{3}$$

It is obvious that any measures smaller and larger than λ will result in negative and positive terms, respectively, wherein the tradeoff λ is analogous to a threshold. Therefore, the closed-form solution $\mathbf{p}$, which minimizes Q by preserving all the non-positive measures, is obtained as:

$$p_i = \begin{cases} 1, & \text{if } C_i \leq \lambda \\ 0. & \text{otherwise} \end{cases}. \tag{4}$$

The final correspondence is obtained as $\boldsymbol{\mathcal{I}}^* = \{(\mathbf{x}_i, \mathbf{y}_i) | p_i = 1, i = 1, 2, ..., N\}$. Please refer to Section III-D for the post-treatment of $\boldsymbol{\mathcal{I}}^*$ called inlier retrieval.

B. Context-aware Locality Measure (CALM)

Many feature extractors, e.g., SIFT [2], SURF [40], etc., implement feature extraction based on the scale and orientation invariance. The scale and orientation infused with the extracted features have strong consistency among inliers. For instance, to an inlier pair, their ratio of scales mostly corresponds to the scale factor between the two image regions [49], and their orientation differences should also be similar to those of other inlier pairs. These observations inspire us to interpret the consistency from scale, orientation, etc., to our dissimilarity measure. Therefore, the context-aware locality measure (CALM) is designed. It measures the dissimilarity based on the neighborhood relationship, inter-neighborhood distance, and context information.

[1] $\boldsymbol{\mathcal{S}}$ and $\boldsymbol{\mathcal{I}}_0$: the universal correspondence set and putative inlier set extracted from an image pair using the loosest NNDR threshold τ_0 and τ, respectively. All the pairs involving repetitive features are removed as a pretreatment guaranteeing $\boldsymbol{\mathcal{I}}_0 \subset \boldsymbol{\mathcal{S}}$. $\boldsymbol{\mathcal{I}}$: the inlier pool, which is the currently known correspondences consisting of inliers. $\boldsymbol{\mathcal{C}}$: the candidate pool obtained by a set difference operation $\boldsymbol{\mathcal{S}} \backslash \boldsymbol{\mathcal{I}}$. The inlier pool will be gradually enriched by a subset of the candidate pool that is preserved by the dissimilarity measure.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2961480, IEEE Transactions on Image Processing

The neighborhood relationship desires the one-to-one matching among the corresponding neighboring inliers given a point pair. Ideally, the perfect match achieves when every neighbor from $\boldsymbol{\mathcal{N}}_{\hat{\mathbf{x}}_i}$ is incident to exactly one neighbor with the same index from $\boldsymbol{\mathcal{N}}_{\mathbf{y}_i}$. It is formulated as:

$$g_i = \frac{1}{2K}\left(\sum_{j|\mathbf{y}_j \in \boldsymbol{\mathcal{N}}_{\mathbf{y}_i}} b(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) + \sum_{j|\hat{\mathbf{x}}_j \in \boldsymbol{\mathcal{N}}_{\hat{\mathbf{x}}_i}} b(\mathbf{y}_i, \mathbf{y}_j)\right), \quad (5)$$

where the binary distance $b(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) = 1$, if $\hat{\mathbf{x}}_{j|\mathbf{y}_j \in \boldsymbol{\mathcal{N}}_{\mathbf{y}_i}} \notin \boldsymbol{\mathcal{N}}_{\hat{\mathbf{x}}_i}$, or $b(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) = 0$, otherwise. The same rule is applied to $b(\mathbf{y}_i, \mathbf{y}_j)$. For one point $\mathbf{x}_i$ from a point set $\mathbf{X}$, its indexes j for summation are determined by the neighbors $\mathbf{y}_j$ of the paired point $\mathbf{y}_i$ from another set $\mathbf{Y}$. The normalization term $1/2K$ ensures that the range of g_i is $[0, 1]$.

The inter-neighborhood distance penalizes the pairs that have similar neighborhood relationships yet quite different relative distribution. Such inconsistency usually occurs when the overlapping area of an image pair only governs a small proportion of either one of the images (e.g., under planar scaling and translation). Given a $K \times 2$ neighborhood set $\boldsymbol{\mathcal{N}}_{\mathbf{z}_i}$, we define the $2K \times 1$ vector $d_{\mathbf{z}_i} = vec(\boldsymbol{\mathcal{N}}_{\mathbf{z}_i}) - \overline{\boldsymbol{\mathcal{N}}_{\mathbf{z}_i}} \cdot \mathbf{1}$, where $vec(\cdot)$ denotes the row-wise vectorization of a matrix, $\mathbf{1}$ denotes a column vector of all ones, and the scalar $\overline{(\cdot)}$ denotes the mean of a matrix. We can therefore write the inter-neighborhood distance as:

$$a_i = 1 - \frac{d_{\hat{\mathbf{x}}_i}^T d_{\mathbf{y}_i}}{\sqrt{d_{\hat{\mathbf{x}}_i}^T d_{\hat{\mathbf{x}}_i} d_{\mathbf{y}_i}^T d_{\mathbf{y}_i}}}, \quad (6)$$

where the second term is basically the 2-D correlation coefficient.

The context information is parameterized using a local histogram, encouraging pairs that have small chi-square distance. The histogram greatly related to the reference angular coordinate echoes our previous concern about orientation. Given a point $\hat{\mathbf{x}}_i$ and the relative coordinates of its K neighboring inliers, i.e., $\hat{\mathbf{x}}_{j|\hat{\mathbf{x}}_j \in \boldsymbol{\mathcal{N}}_{\hat{\mathbf{x}}_i}} - \hat{\mathbf{x}}_i$, a local histogram $\mathbf{h}^{\hat{\mathbf{x}}_i} = \{h_r^{\hat{\mathbf{x}}_i} | r = 1, 2, ...R\}$ centered at $\hat{\mathbf{x}}_i$, with R bins that are uniform in log-polar space [50], can be established to extract the context information of $\hat{\mathbf{x}}_i$ as:

$$h_r^{\hat{\mathbf{x}}_i} = |\hat{\mathbf{x}}_j : (\hat{\mathbf{x}}_j - \hat{\mathbf{x}}_i) \in bin(r), \hat{\mathbf{x}}_j \in \boldsymbol{\mathcal{N}}_{\hat{\mathbf{x}}_i}|. \quad (7)$$

By applying the same histogram to $\mathbf{y}_i$, the context dissimilarity ψ_i of the i-th point pair can be quantified using the chi-square distance

$$q_i = \frac{1}{2}\sum_{r=1}^{R} \frac{(h_r^{\hat{\mathbf{x}}_i} - h_r^{\mathbf{y}_i})^2}{h_r^{\hat{\mathbf{x}}_i} + h_r^{\mathbf{y}_i}}, \quad (8)$$

within a fixed range. More specifically, for any point pairs $(\hat{\mathbf{x}}_i, \mathbf{y}_i)$, we have $q_i \in [0, K]$.

The CALM is then formulated according to the following intuitions. The neighborhood relationship $\mathbf{g}$ exploits only the binary relationship within the neighborhood. It may blindly produce zero cost once the neighborhood indexes are identical, without considering the actual spatial consistency. The local context information $\mathbf{q}$ is the most informative component since it considers both the orientation and distance. However, it may produce a high cost if the reference direction is biased. Hopefully, the summation $(\mathbf{g}+\mathbf{a})$ can regularize $\mathbf{g}$ by the inter-neighborhood distance $\mathbf{a}$, because $\mathbf{a}$ is rarely zero in practice. And the multiplication of $\mathbf{q}$ and $(\mathbf{g}+\mathbf{a})$ can alleviate the false large cost from $\mathbf{q}$ itself. Therefore, we formulate the CALM by incorporating Eq. 5, Eq. 6 and Eq. 8 as:

$$C^{\boldsymbol{\mathcal{N}} \subset \boldsymbol{\mathcal{I}}}(\hat{\mathbf{x}}_i, \mathbf{y}_i) = q_i \left(g_i + \omega a_i\right), \quad (9)$$

where the constant $\omega \geq 0$ controls the strength of the regularization. Our CALM can yield reasonable measures for various scenarios accordingly, as displayed in Fig. 2. We find that the form of Eq. 9 can yield better performance when compared against a typical form, i.e., $q_i + \omega_1 g_i + \omega_2 a_i$. The explanation could be that there is a non-linear interaction between $\mathbf{q}$ and other criteria. The setting of bin and neighboring inlier number for our CALM is investigated in Section IV.

(a) (b) (c) (d)

Figure 2: The illustration of the four types of distributions that may appear during the measure. The red and blue denote the two feature point sets, with the features that are being measured and their neighboring inliers denoted by blank and filled ones, respectively. (a): in the beginning, the two sets may have a scale difference. Since the radius of CALM is determined by the K-th neighboring inlier, the relative distribution captured should be similar. (b): after the pre-alignment, the neighboring inliers given a candidate pair should be already roughly aligned. If the central features being measured are inliers, they should also be aligned as well. Otherwise, they will have large variations in the respective neighborhood caused by the spatial deviation of the central points. (c): by applying the reference angular coordinate and regularized neighborhood relationship, the high cost caused by the angular bias can be alleviated, which makes the measure reasonable. (d): when the feature pair has the same neighborhoods yet large deviation (which is typical for features at the fringe of the overlapping area), the estimation using only neighborhood relationships will yield the false positive, and the falsely aligned pairs may further result in unexpected distortion. In contrast, our CALM can return a large dissimilarity measure thanks to the local context information and inter-neighborhood distance.

C. Candidate Bifurcation

The candidate bifurcation intends to reduce the distance and structural dissimilarity of the potential inliers, while making outlier pairs more conspicuous for pruning. To this end, the approximate thin-plane spline (ATPS) transformation [51] parameterizing the transformation $\mathbf{f}$ is employed. Given the inlier set $\boldsymbol{\mathcal{I}} = \{(\mathbf{x}_i', \mathbf{y}_i')\}_{i=1}^{N'}$, the universal set $\boldsymbol{\mathcal{S}} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ and the radial basis function $U(u) = u^2 \log u$, the unknown coefficient $\boldsymbol{\theta}_{(N'+3)\times 2}$ is found by solving the linear system:

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\mathcal{K}}' + \eta \mathbf{I} & \mathbf{H}' \\ \mathbf{H}'^T & \mathbf{O} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{Y}' \\ \mathbf{0} \end{pmatrix}, \quad (10)$$

where $\boldsymbol{\mathcal{K}}'_{N' \times N'}$ is the radial basis kernel with each entry $\boldsymbol{\mathcal{K}}'_{ij} = U(\|\mathbf{x}_i' - \mathbf{x}_j'\|)$, $\mathbf{H}' = (1, \mathbf{X}')$ is the $N' \times 3$ homogeneous coordinate, $\mathbf{O}_{3\times 3}$ and $\mathbf{0}_{3\times 2}$ are terms of all zeros, and the identity matrix $\mathbf{I}_{N' \times N'}$ is for regularization with the constant

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2961480, IEEE Transactions on Image Processing

η controlling the weight. Subsequent to which the updated source point set is obtained by $\hat{\mathbf{X}} = \begin{pmatrix}\boldsymbol{\mathcal{K}} & \mathbf{H}\end{pmatrix}\boldsymbol{\theta}$, where $\boldsymbol{\mathcal{K}}$ is the $N \times N'$ basis containing $\boldsymbol{\mathcal{K}}_{ij} = U(\|\mathbf{x}_i - \mathbf{x}'_j\|)$, and $\mathbf{H} = (1, \mathbf{X})$ is the $N \times 3$ homogeneous coordinate.

D. Main Process

Our SIR consists of three major steps, which are the initialization, stepwise process, and inlier retrieval. During the initialization, the normalized[2] universal set $\boldsymbol{\mathcal{S}}$ and its subset, the putative inlier set $\boldsymbol{\mathcal{I}}_0$, are obtained with the loosest τ_0 and strict τ, respectively. The CALM then filters $\boldsymbol{\mathcal{I}}_0$ to establish the pre-alignment on $\boldsymbol{\mathcal{S}}$. After which the inlier and candidate pools are obtained by the filtering $\boldsymbol{\mathcal{I}} \xleftarrow{\text{CALM}} \boldsymbol{\mathcal{I}}_0$ and set difference $\boldsymbol{\mathcal{C}} \leftarrow \boldsymbol{\mathcal{S}} \backslash \boldsymbol{\mathcal{I}}$, respectively.

The stepwise process involves the alternating inlier pool enrichment and candidate pool bifurcation. In each iteration, a subset $\boldsymbol{\mathcal{C}}'_m$ of the candidate pool, with its cardinality equaling current $|\boldsymbol{\mathcal{I}}|$, is selected in the order of index. The selected candidates undergo the pruning, which is up to what distance ϵ are they considered as egregious outlier pairs, preserving $\boldsymbol{\mathcal{C}}_m = \{(\hat{\mathbf{x}}_i, \mathbf{y}_i) | \|\hat{\mathbf{x}}_i - \mathbf{y}_i\| \leq \epsilon, (\hat{\mathbf{x}}_i, \mathbf{y}_i) \in \boldsymbol{\mathcal{C}}'_m\}_{i=1}^{N''}$. The preserved candidates are then judged by CALM yielding the intermediate inlier set $\boldsymbol{\mathcal{I}}_m$. Hence the inlier pool is enriched by $\boldsymbol{\mathcal{I}} \leftarrow \boldsymbol{\mathcal{I}} \cup \boldsymbol{\mathcal{I}}_m$. After which the transformation $\mathbf{f}$ is built to align the pairs from $\boldsymbol{\mathcal{I}}$ exclusively. Meanwhile, it also extrapolates other pairs accordingly so that the similarity for the potential inlier and outlier pairs is bifurcated. Our SIR iterates until the candidate pool is traversed.

The inlier retrieval maximizes the inlier pool based on the abundant and well-distributed inliers. It first reconstructs the candidate pool again by $\boldsymbol{\mathcal{C}} \leftarrow \boldsymbol{\mathcal{S}} \backslash \boldsymbol{\mathcal{I}}$, and then carries out the enrichment and bifurcation over the whole $\boldsymbol{\mathcal{C}}$ to retrieve falsely rejected inlier pairs. The final inlier set $\boldsymbol{\mathcal{I}}^*$ and image transformation coefficient $\boldsymbol{\theta}^*$ are yielded after the retrieval, where the latter can further transform the sensed image to overlay it over the reference one by $I^o = \mathbf{f}(I^s, \boldsymbol{\theta}^*)$ generating the transformed image [1]. The pseudo-code is provided in Algorithm 1.

Algorithm 1: Stepwise Image Registration

input : Two images I^s and I^r
output : Transformed image I^o, final inlier set $\boldsymbol{\mathcal{I}}^*$

1 Establish the normalized $\boldsymbol{\mathcal{S}}$ and its subset $\boldsymbol{\mathcal{I}}_0$ under threshold τ_0 and τ, respectively;
2 Filter the putative inliers by $\boldsymbol{\mathcal{I}}_0 \xleftarrow{\text{CALM}} \boldsymbol{\mathcal{I}}_0$;
3 Pre-align $\boldsymbol{\mathcal{S}}$ by the mapping $\mathbf{f}$ from $\boldsymbol{\mathcal{I}}_0$;
4 Initialize the inlier pool by $\boldsymbol{\mathcal{I}} \xleftarrow{\text{CALM}} \boldsymbol{\mathcal{I}}_0$;
5 Initialize the candidate pool by $\boldsymbol{\mathcal{C}} \leftarrow \boldsymbol{\mathcal{S}} \backslash \boldsymbol{\mathcal{I}}$;
6 **while** $\boldsymbol{\mathcal{C}}$ *is not traversed* **do**
7 Select the candidates $\boldsymbol{\mathcal{C}}'_m$ where $|\boldsymbol{\mathcal{C}}'_m| = |\boldsymbol{\mathcal{I}}|$;
8 Prune the selected candidates by $\boldsymbol{\mathcal{C}}_m \xleftarrow{\epsilon} \boldsymbol{\mathcal{C}}'_m$;
9 Determine the new inlier set by $\boldsymbol{\mathcal{I}}_m \xleftarrow{\text{CALM}} \boldsymbol{\mathcal{C}}_m$;
10 Enrich the inlier pool by $\boldsymbol{\mathcal{I}} \leftarrow \boldsymbol{\mathcal{I}} \cup \boldsymbol{\mathcal{I}}_m$;
11 Refine the coefficient $\boldsymbol{\theta}$ based on enriched $\boldsymbol{\mathcal{I}}$;
12 **end**
13 Re-construct the candidate pool by $\boldsymbol{\mathcal{C}}' \leftarrow \boldsymbol{\mathcal{S}} \backslash \boldsymbol{\mathcal{I}}$;
14 Prune the selected candidates by $\boldsymbol{\mathcal{C}} \xleftarrow{\epsilon} \boldsymbol{\mathcal{C}}'$;
15 Retrieve remaining inliers by $\boldsymbol{\mathcal{I}}_{M+1} \xleftarrow{\text{CALM}} \boldsymbol{\mathcal{C}}$;
16 Obtain the final inlier set by $\boldsymbol{\mathcal{I}}^* \leftarrow \boldsymbol{\mathcal{I}} \cup \boldsymbol{\mathcal{I}}_{M+1}$;
17 Obtain the final coefficient $\boldsymbol{\theta}^*$ and transform I^s;

IV. Analysis and Implementation Details

The bin density (BD) of CALM, number of neighbors K for neighborhood construction, and NNDR threshold τ for establishing the putative inlier set play cruel roles in our SIR. First, the BD represents the strictness of the dissimilarity measure on the context information. A very dense BD, say $(50, 120)$ for radial and tangential directions, may yield a high cost due to the slight bias on the reference angle. Second, K determines how many samples are used to yield the CALM. A large K leads to a relatively large neighborhood and will involve more inconsistency. Also, too high a K (e.g., $K \approx |\boldsymbol{\mathcal{I}}_0|$) may ruin the registration if the initial inlier pool is quite limited. Third, the filtering and pre-alignment used in Line 2 and 3 of Algorithm 1 can attenuate the negative effect of the initial outliers in $\boldsymbol{\mathcal{I}}_0$, which suggests that our SIR is not sensitive to the τ. Though using a strict τ can definitely lower the outlier ratio and cardinality of $\boldsymbol{\mathcal{I}}_0$ and thus achieves a better accuracy-efficiency tradeoff, the correspondences established at this case may not be sufficient to perform the stepwise process nor could it guide a fine image transformation, particularly when the image pair suffers low resolution, low overlap ratio or large viewpoint variations. Fourth, the threshold τ also influences the performance of our SIR when the images are contaminated by noises, e.g., the Gaussian noise, at which case the correctness of $\boldsymbol{\mathcal{I}}_0$ might be further lowered due to the random deviation applied to each image intensity.

According to the above prior knowledge, the optimal parameter setting of our SIR is investigated as follows. The CALM threshold λ and pruning threshold ϵ are first evaluated by fixing the inlier ratio (IR), inlier number (IN), BD, K, and τ to 0.1, 300, $(5, 12)$, 5, and 1.25, respectively. After which the previously fixed parameters are evaluated under the optimal λ and ϵ. To synthesize the feature sets satisfying the requirements of number and ratio, the following technique is introduced. The inlier pairs are randomly selected from the manually confirmed ground-truths of an image pair. The random outliers within the image boundary (i.e., using the Matlab *rand* function) along with their corresponding SIFT descriptors are generated using the VLFEAT toolbox [3] (i.e., using the *vl_sift* function). Finally, the random outlier pairs are concatenated to the inlier pairs to meet the ratio requirement. For example, the inlier ratio 0.1 can be obtained using 300 inlier pairs and 2700 outlier pairs. To apply the Gaussian noise to images [3], the standard deviation of intensity given an image

[2]The normalization process rescales the feature point sets, generalizing arbitrary feature sets to have zero means and unit variances. The same spatial scale is beneficial to further thresholding. It also roughly overlays the source and target sets, by which the matching ambiguity is reduced [21], [22], [23].

[3]Unless otherwise stated, the noise is always applied to the source image I^s in the remainder of this paper.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2961480, IEEE Transactions on Image Processing

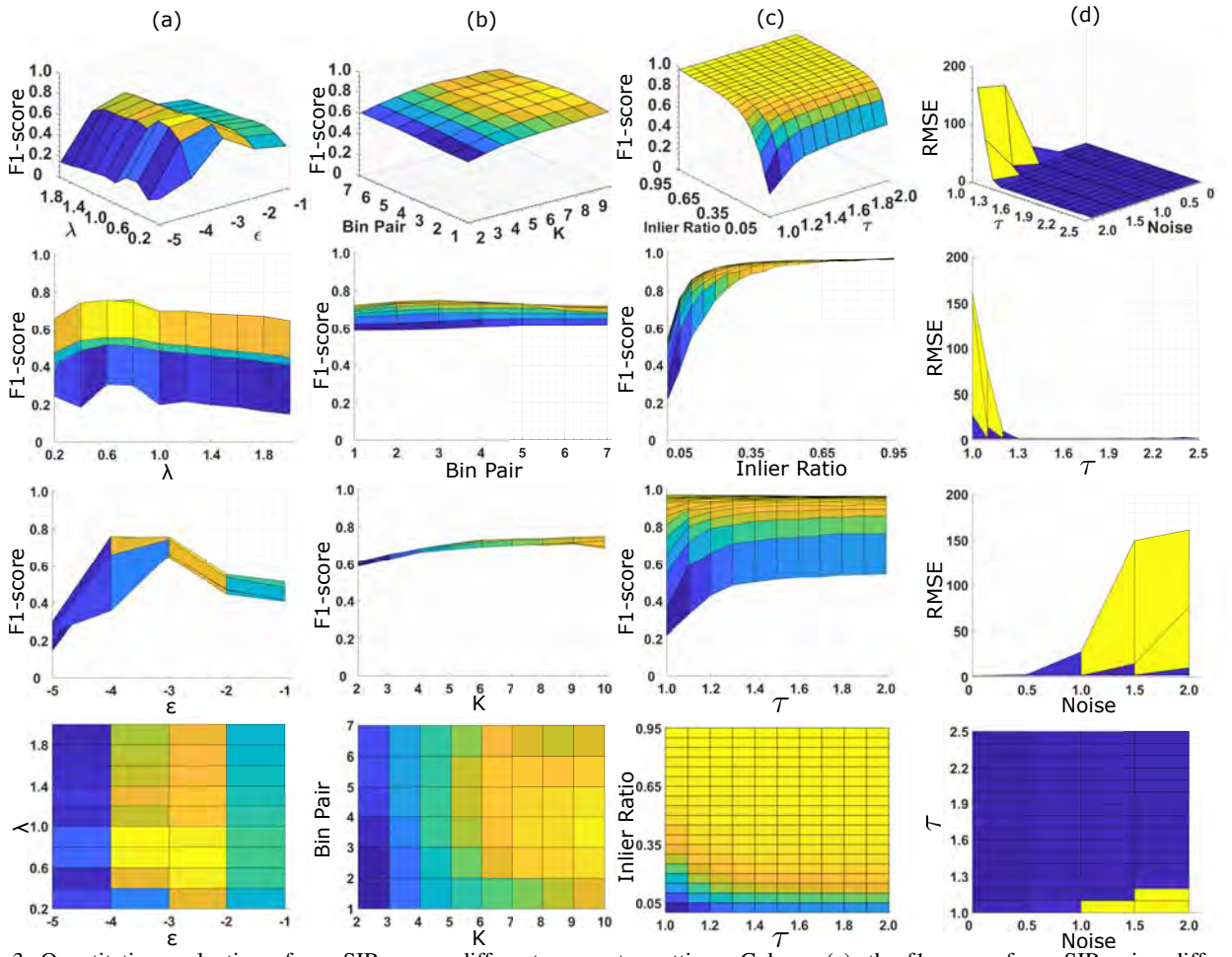

Figure 3: Quantitative evaluation of our SIR across different parameter settings. Column (a): the f1-score of our SIR using different λ-ϵ settings. These two parameters determine to what extent could a correspondence be rejected by the CALM and pruning, respectively. The axis for ϵ is spaced logarithmically, e.g., -5 denotes $1e-5$. Column (b): the f1-score of our SIR using different neighborhood settings. The bin pair and K define the density and locality of the CALM, respectively. Ten pairs, i.e., $\{(3,8),(4,10),(5,12),(8,20),(10,24),(15,36),(20,48)\}$, are labeled as $\{1,2,\cdots,7\}$ on the axis, respectively. Column (c): the f1-score of our SIR with varied strictness of putative set against the correctness of the universal set. Column (d): the RMSE of our SIR with different strictness of putative set against the Gaussian noise. Five noise levels, i.e., $0, 50\%, 100\%, \ldots, 200\%$ are labeled as $0, 0.5, 1.0, \ldots, 2.0$. For each column, views from xyz-axis, xz-axis, yz-axis, and xy-axis are used. Totally 100 replications are conducted for each treatment.

is firstly measured, the Gaussian noise which is in proportion to the intensity deviation is then randomly generated using the Matlab *imnoise* function. Finally, the noise grid is added to the image intensity to realize the noise contamination. 100 replications are conducted for each treatment. The results are shown in Fig. 3.

For the pair of λ and ϵ, our SIR achieves the optimal performance at $\lambda = 0.8$ and $\epsilon = 0.0001$. For the BD and K, our SIR is favored by the treatments with sparse histogram and larger K, and within an acceptable scope of performance, say f1-score ≥ 0.7, different treatments have no obvious variations on f1-score. For the NNDR threshold τ against the outliers, we can see that the f1-scores are not sensitive with respect to τ given IR$\geq$ 0.15, by which the effect of filtering and pre-alignment is demonstrated. The insensitivity implies that our SIR can be generalized to different local descriptors [52]. For the NNDR threshold τ against the noise, the RMSE is acceptable for all the noise levels when $\tau \geq 1.3$. Overall, the default parameter setting of our SIR is determined as $\lambda = 1.2$, $\epsilon = 0.001$, BD$= (5, 12)$, $K = 5$ and $\tau = 1.3$, according to the results from Fig. 3 and trial-and-error on the real data. The ATPS smoothness is set to $\eta = 0.5$ empirically.

The ablation study is then conducted to investigate to what extent can the pruning, retrieval, and stepwise process affect our SIR in terms of robustness and efficiency. The non-pruned (SIR-NP), non-retrieval (SIR-NR), and non-stepwise (SIR-One) variants of our SIR are investigated. For SIR-NP, the operations of Line 8 and 14 from Algorithm 1 are removed. For SIR-NR, the operations of Line 13, 14, and 15 from Algorithm 1 are removed. And for SIR-One, the final inlier set is directly obtained by $\boldsymbol{\mathcal{I}}^* \xleftarrow{\text{CALM}} \boldsymbol{\mathcal{S}}$ after the pre-alignment of Line 3 Algorithm 1. Three factors: (i) the IN of universal set $\boldsymbol{\mathcal{S}}$, (ii) the IR of $\boldsymbol{\mathcal{S}}$, and (iii) the IR of putative inlier set $\boldsymbol{\mathcal{I}}_0$, i.e., the seed correctness (SC), are selected for quantitative evaluation. Three scenarios are therefore designed, each of which varies one factor with the other two being fixed, as shown in Table I. GLPM [15] is chosen as the baseline for comparison since it also requires $\boldsymbol{\mathcal{S}}$ and $\boldsymbol{\mathcal{I}}_0$ as the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2961480, IEEE Transactions on Image Processing

input yet without our stepwise process, CALM and candidate bifurcation. The f1-score is used as the evaluation criterion. It estimates the balance between the recall and precision. The experimental results are shown in Fig. 4.

Table I: Three experimental scenarios for quantitative evaluation of the stepwise manner. S: scenario; IR: inlier ratio; IN: inlier number. SC: seed correctness.

S	IR of $\mathcal{S}$	IN of $\mathcal{S}$	SC	$\lvert\mathcal{I}_0\rvert$
1	$0.05, 0.10, \ldots, 1.0$	300	1.0	15
2	0.1	$50, 80, \ldots, 320$	1.0	15
3	0.1	300	$1.0, 0.9, \ldots, 0$	50

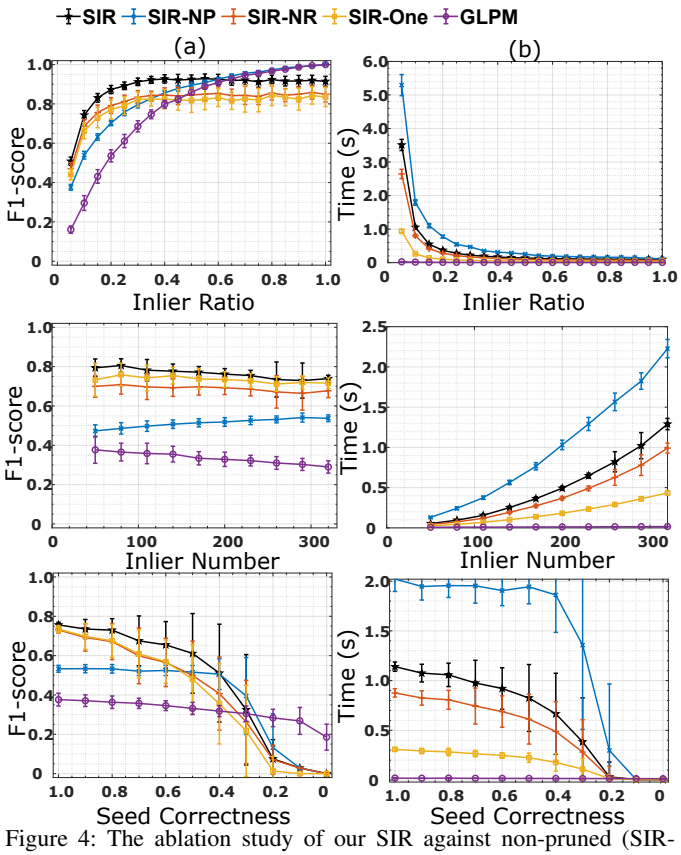

Figure 4: The ablation study of our SIR against non-pruned (SIR-NP), non-retrieved (SIR-NR), one-step variants (SIR-One) and the baseline method GLPM, respectively. Column (a) and Column (b) are the results on f1-score and time cost, respectively. Rows from top to bottom corresponds to Scenario 1, 2, and 3, respectively. The error bars indicate the standard deviations of the errors. 100 replications are conducted for each treatment. Best viewed in color.

In Scenario 1 the f1-score of our SIR outperforms the variants and baseline when IR$\leq$ 0.6, after which the SIR-NP and GLPM achieves a larger f1-score. It suggests that the pruning and ATPS can reduce the ambiguity caused by low IR, they however backfire when IR is high. The limited putative inlier set, i.e., $|\mathcal{I}_0| = 15$, is insufficient for structural description since many distant points may have the same neighboring inliers yet quite different relative distribution. The stepwise manner enriching the inlier pool gradually alleviates this issue. The reason why our SIR cannot achieve f1-score 1.0 when IR$=1$ is mainly because of the drawback of using a radial basis function as the general image deformation model. The ATPS at the area with sparse control points may misalign

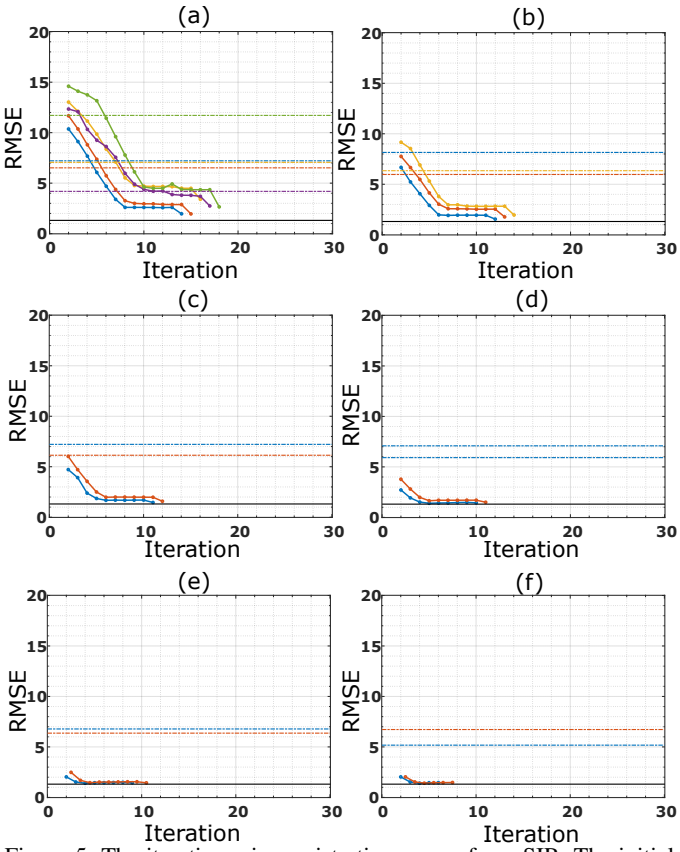

Figure 5: The iteration-wise registration error of our SIR. The initial inlier set is manually created by randomly selecting 15, 25, 35, 55, 105, and *all* pairs from $\mathcal{I}_0$, as shown in (a) to (f), respectively. The marked lines show the average iteration-wise error at the i-th step. Various colors are used to group trials ended up with the same iteration counts, and the dashed lines with the same colors indicate the associated final RMSE from GLPM. The black line on the bottom indicates the systematic error.

the true inlier pairs, leading to a false rejection by the pruning. Fortunately, such drawback is tolerable as our interest is to recover inliers from the lowest possible NNDR threshold (i.e., the IR is usually $<$ 0.4). In Scenario 2, as the potential inlier grows, the difficulty of using $|\mathcal{I}_0| = 15$ to recover the remaining inlier pairs also increases. As expected, the stepwise manner gives the best and steady f1-score, and its advantage against others is obvious. Whereas GLPM yields unacceptable results due to the limited $|\mathcal{I}_0|$ and low inlier ratio 0.1. The low f1-score, obtained by the non-pruned version results from the imbalance of high recall and low precision, reflects not only the availability of the pruning in matching ambiguity reduction, but also the sensitivity of the CALM towards point deviation. Further investigation and discussion about the sensitivity are provided in Secion VI. Scenario 3 demonstrates that the stepwise manner enhances the tolerance of our SIR towards the falseness of $\mathcal{I}_0$, as the apparent advantage can be seen when the inlier ratio of $\mathcal{I}_0$ is $\leq$ 0.5, after which the results are deteriorated. When it comes to efficiency, we see that SIR-NP is the most time-consuming, which requires nearly 5.5s to process 6K feature pairs. Thus it validates the availability of the pruning on efficiency improvement. GLPM features perfect time-saving result since it does not

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2961480, IEEE Transactions on Image Processing

involve radial basis function based transformation. Overall, the stepwise manner can greatly improve the accuracy and robustness at the cost of tractable time.

We also experiment to reveal the iteration-wise behavior of our SIR with regard to the sufficiency of the initial inlier set. In this experiment we set IN$=300$, IR$=0.2$, and $\tau = 1.3$, and manually select 15, 25, 35, 55, 105, and all pairs from $\boldsymbol{\mathcal{I}}_0$. Totally 100 landmark pairs are manually labeled to compute the RMSE. The GLPM is chosen as the baseline method. Over the 100 replications, the representative trials ended up with the same iteration counts are selected to calculate the average iteration-wise RMSE and baseline RMSE. The results are shown in Fig. 5. The line graphs suggest that the iteration counts are inversely proportional to the cardinality of the initial inlier set. When $|\boldsymbol{\mathcal{I}}_0|$ is small, e.g., ≤ 35, obvious RMSE drops for the first half of each line can be observed, while for the rest part, the RMSEs tend to be steady. Particularly, for Fig. 5a, i.e., when $|\boldsymbol{\mathcal{I}}_0| = 15$, the RMSEs significantly drop at the last iteration due to the employment of the retrieval process. As $|\boldsymbol{\mathcal{I}}_0|$ increases, the overall RMSE approaches the systematic error at an earlier iteration, and the improvement at the last iteration becomes trivial. The explanation for the observed phenomenon lies in the sufficiency of the inlier pool (which plays the role of control points) for ATPS transformation establishment. A small set of control points clustered regionally may fail to present the image deformation far apart, and therefore incur misalignment and false rejection of inliers. As the control points are gradually enriched by the candidates, the ATPS transformation becomes finer in a larger area until reaching the saturation. The retrieval process can reexamine the candidates based on the finest alignment and CALM, and recovers inliers that are previously falsely rejected. The latter, if positioned in a region with sparse control points, might greatly decrease the RMSE in a global manner.

The implementation details are provided based on Algorithm 1. First of all, given the universal set $\boldsymbol{\mathcal{S}}$, two lists storing the indexes of $\boldsymbol{\mathcal{I}}_0$ and the rest feature pairs are generated after Line 1. Throughout our SIR the first list will be enriched to yield the final inliers, and the second list will be traversed, as stated in Line 16 and Line 6, respectively. The first use of CALM in Line 2 differs from the rest, as its judgment is based on the initial features before the pre-alignment. At which time, the reference angular coordinate is initiated by the orientation from the SIFT descriptor. Meanwhile $\omega = 0$ from Eq. 9 is used for ensuring the scaling, translation, and rotation invariance of CALM. Whereafter the reference angular coordinate is set as the direction from the point currently being measured to its nearest neighbor, and $\omega = 1$ is adopted to impose the equal priority between neighborhood relationship and inter-neighborhood distance. The construction in Line 4 improves the correctness of the inlier pool since the matching ambiguity is reduced by the pre-alignment. In Line 7 the requirement of equal cardinality instead of a fixed step size addresses the possible dilemma caused by an extremely unbalanced inlier-to-candidate ratio so that the capability of CALM in the stepwise process will not be exceeded or wasted. Such adaptivity also implies that the actual iteration number M for different image pairs is flexible. The pruning in Line 14 ensures that the outliers, which are not consistent with the ATPS transformation, are removed. This helps the following retrieval process safely recover the false negatives to maximize the inlier pool as well as minimize the registration error. When the size of the inlier pool is small, this step might greatly decrease the registration error in a global manner.

The KNN searching has $\mathcal{O}((N+K)\log(N))$ complexity using the k-d tree [57]. In addition to which the dissimilarity measure using CALM is of $\mathcal{O}(KN)$ complexity. The matrix inversion of the linear system Eq. 10 with $\mathcal{O}(N^3)$ complexity is the computational bottleneck of our SIR. Fortunately, by employing ATPS the actual complexity is reduced to $\mathcal{O}(N'^3)$. When N is very large (e.g., $N' = 0.1N$, which is analogous to inlier ratio 0.1), the saving factors in processing time could be 1000x [51]. Overall, our SIR has $\mathcal{O}(N\log(N) + N'^3)$ complexity, since $K \ll N$.

V. Experiment

To test the performance and generality of our SIR, extensive experiments including feature matching, image registration, and image retrieval are carried out. Eight state-of-the-art methods, namely RANSAC [4], PROSAC [6], LPM [14], GLPM [15], ICF [16], GS [11], CPD [22] and GLMD [23] are used for comparison.

The experiments are performed on a laptop with a 2.6 GHz Intel Core CPU and 16 GB memory using Matlab code. The open-source VLFEAT toolbox [3] is employed for determining the universal set $\boldsymbol{\mathcal{S}}$ [2] and K-D tree-based neighboring inlier searching. In particular, all the experiments use the loosest NNDR threshold $\tau_0 = 1.0$ to determine the universal set $\boldsymbol{\mathcal{S}}$, and use $\tau = 1.3$ to determine the putative inlier set $\boldsymbol{\mathcal{I}}_0$ for our SIR and GLPM [15]. The overview of our experimental design and the references to the datasets are shown in Table II.

A. Ground-truth and Criterion

Generally, three types of ground-truth are used. For feature matching experiments, its ground-truth is an $N' \times 1$ column vector indicating the indexes of inlier pairs extracted by the SIFT algorithm, where N' denotes the number of inlier pairs out of N feature pairs. The recall, precision, and f1-score are employed as the criteria. An ideal method should be able to yield indexes that are identical to the ground-truth given an image pair so that the three criteria all equal 1. For image registration experiments, the ground-truths are two $M \times 2$ matrices indicating the corresponding coordinates of the human-labeled points, where M denotes the number of labels. The root mean square error (RMSE), and the associated area under curve (AUC) are used as the criteria. An ideal method should be able to overlay the two coordinate sets with RMSE$=0$ and AUC$=1$ given an image pair. For image retrieval experiments, its ground-truth is an $M \times (N-1)$ matrix, where M denotes the number of images in total, N denotes the number of images for each group. The i-th row is the indexes corresponding to the top $N-1$ query results given the i-th query image. The N-score is employed as the criterion. Suppose that $N = 4$, given the i-th query image,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2961480, IEEE Transactions on Image Processing

Table II: The overview of our experimental design. Exp.: experiment. Para. Set.: the parameter setting for the experiment. The changed parameters are listed. Pair Num.: the number of image pairs. Treatment: the number of conditions. Replication: the number of runs for each treatment. FM: feature matching. IReg.: image registration. IRet.: image retrieval. The three synthesized datasets involve the random generation of outlier or noise, therefore 100 replications are conducted.

Exp.	Dataset	Para. Set.	Pair Num.	Treatment	Replication	Result	Method Compared
FM	ACF [49]	Default	40	1	1	Fig. 6, Fig. 8, Table III	RANSAC [4], PROSAC [6], LPM [14], GLPM [15], ICF [16] and GS [11]
	RS [15], [53]	Default	40	1	1	Fig. 6, Fig. 8, Table III	
	Synthesized 1	Default	1	30	100	Fig. 7	
IReg.	RS [15], [53]	Default	40	1	1	Fig. 9, Fig. 13	RANSAC [4], PROSAC [6], LPM [14], GLPM [15], ICF [16], GS [11], CPD [22] and GLMD [23]
	FIRE [54]	Default	134	1	1	Fig. 13, Fig. 14, Table III	
	Synthesized 2	Default	1	1	100	Fig. 10	
	Synthesized 3	Default	1	5	100	Fig. 11	
IRet.	UKBench [55]	$\lambda = 1.6$ is used for IRet.	10000	1	1	Table III	RANSAC [4], PROSAC [6], LPM [14], GLPM [15], ICF [16] and GS [11]
			10000	1	1		
			10000	1	1		
			40000	1	1		
	Holiday [56]		10000	1	1		

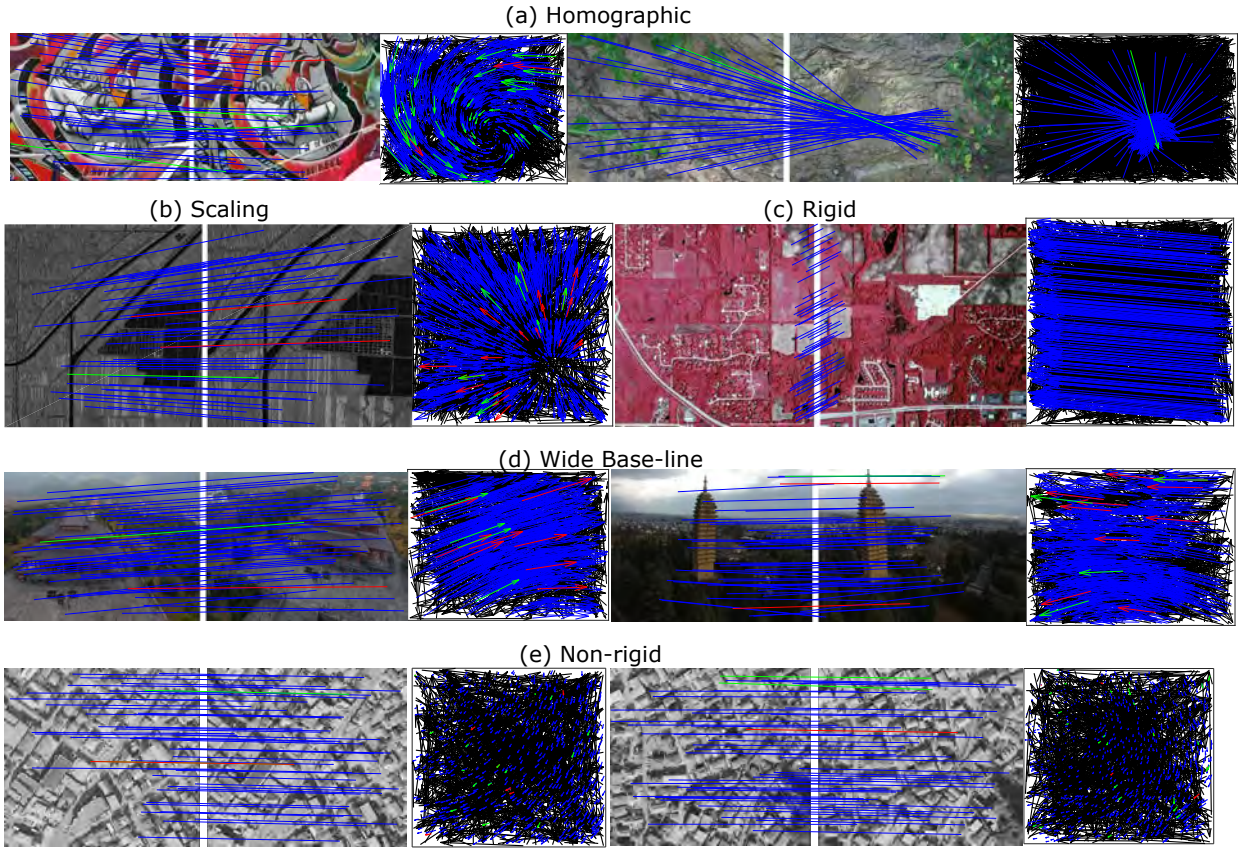

Figure 6: Representative examples of feature matching results on ACF and RS datasets. For each group, the sensed and reference images, and their motion field are displayed. Each arrow in the motion field connects the positions of feature points in image pairs (blue = true positive, black = true negative, green = false negative, red = false positive). For the visual convenience of image pairs, at most 50 randomly selected matches are drawn, and the true negatives are not shown on the sensed and reference image pair. Best viewed in color.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2961480, IEEE Transactions on Image Processing

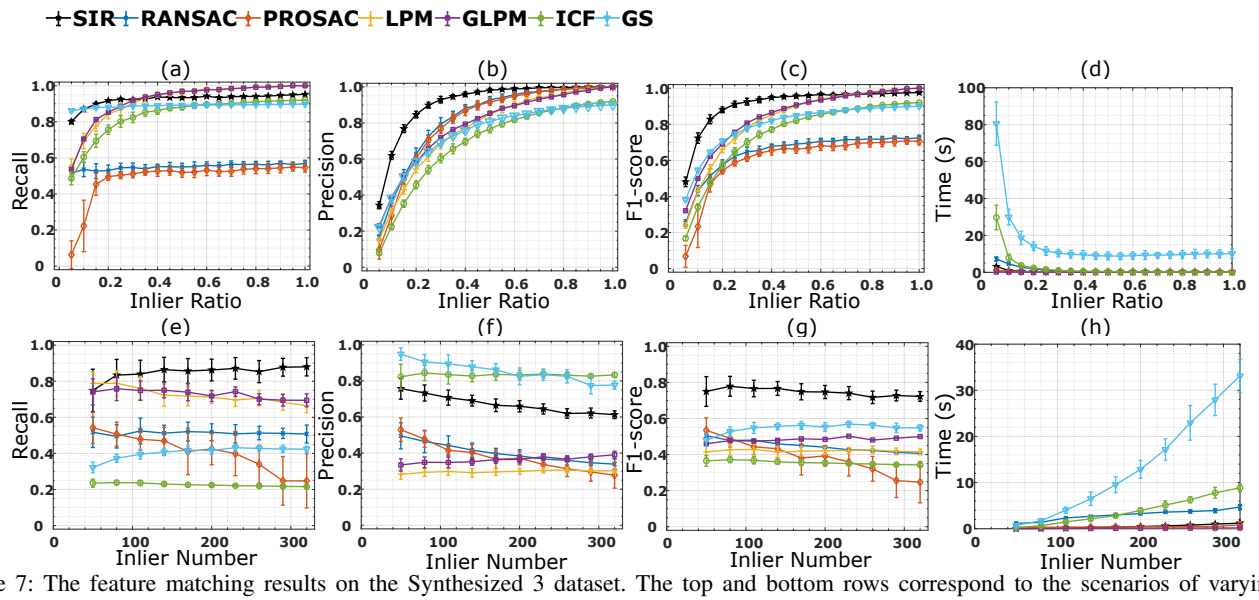

Figure 7: The feature matching results on the Synthesized 3 dataset. The top and bottom rows correspond to the scenarios of varying IR and varying IN, respectively. The last column shows the time cost accordingly. The error bars indicate the standard deviations of the errors. 100 replications are conducted for each treatment. Best viewed in color.

the three remaining images from its group should ideally be contained in the i-th query result.

B. Dataset and Setup

1) Affine covariant features (ACF) dataset: The dataset [49] contains 40 image pairs with changes due to blur, viewpoint, zoom & rotation, light, and JPEG compression. The images are of sizes from 800×640 to 1000×700. For each image pair, its associated homography is provided. The ground-truth is determined by first using the homography to transform the features, and then evaluating the overlap of each feature pair with regards to their SIFT scales [49].

2) Remote sensing (RS) dataset: The dataset [15], [53] contains 40 image pairs captured by high-altitude satellite and small UAVs (SUAVs), respectively. The satellite images feature rigid or affine deformation, while the SUAV images involve non-rigidity due to ground relief variations or imaging viewpoint changes. The images are of sizes from 600×400 to 800×600. The ground-truth for feature matching experiment is obtained by first using our SIR to establish rough correspondence, and then manually confirming the correctness [14], [15]. The ground-truth for image registration experiment is obtained by manually labeling 20 pairs of points between the image pairs, all of which are well-distributed at easily identified places around the interest areas [53], [48].

3) Fundus image registration (FIRE) dataset: The dataset [54] contains three different categories, forming a total of 134 retinal image pairs. Such characteristics are the degree of overlap between images and the presence/absence of anatomical differences. The images are originally acquired with a Nidek AFC-210 fundus camera with a field of view of $45°$, and are down-sampled to the resolution of 582×582 in our experiment. Ground-truth in the form of corresponding image points and a protocol to evaluate registration accuracy are provided.

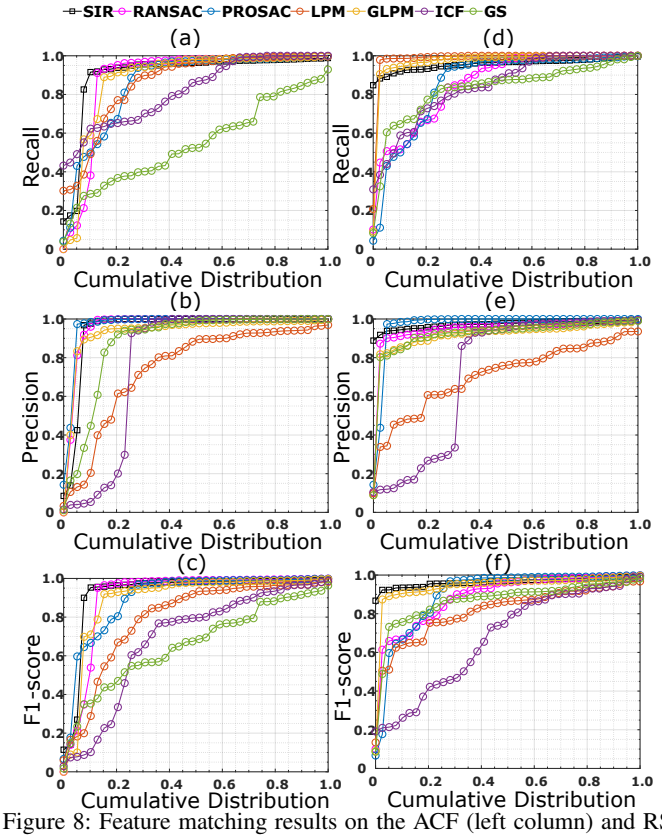

Figure 8: Feature matching results on the ACF (left column) and RS (right column) datasets. Best viewed in color. For the corresponding area under curves, please see Table III.

4) Synthesized 1-3 dataset: They are produced using three types of data augmentation techniques on a randomly chosen image pair from RS dataset. The controllable treatments provide an excellent way to evaluate the performance on accuracy and efficiency quantitatively. For synthesized 1 dataset, Sce-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2961480, IEEE Transactions on Image Processing

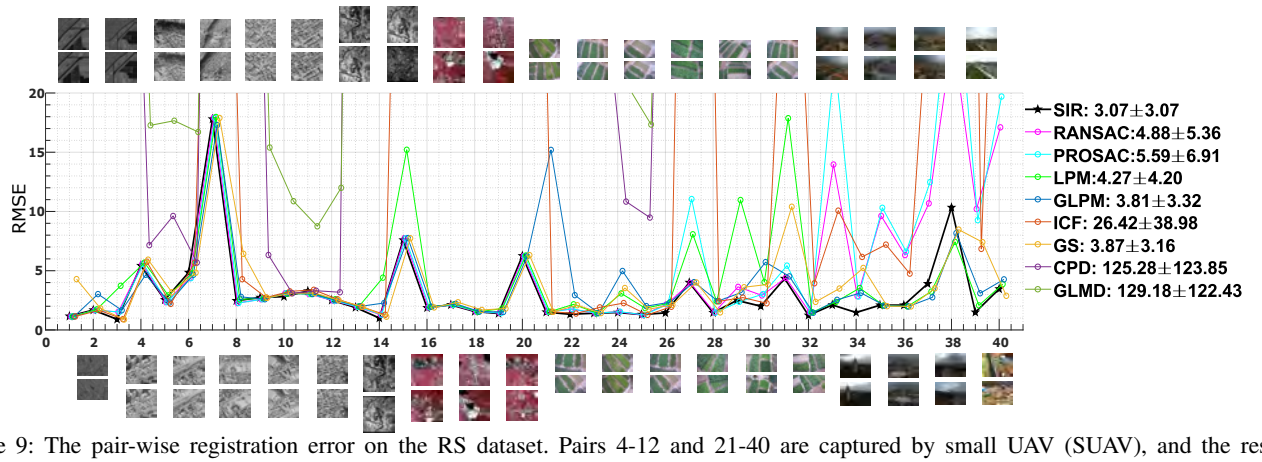

Figure 9: The pair-wise registration error on the RS dataset. Pairs 4-12 and 21-40 are captured by small UAV (SUAV), and the rest are captured by high-altitude satellites. The overall RMSE and standard deviation are indicated in the legend. Best viewed in color.

nario 1 and 2 of Table I are followed to specify the inlier ratio (IR) and inlier number (IN) of the universal set $\boldsymbol{\mathcal{S}}$, resulting in 20 treatments on IR and 10 treatments on IN. For synthesized 2 dataset, the IR and IN of $\boldsymbol{\mathcal{S}}$ are fixed as 0.2 and 300, respectively, resulting in 1 treatment. For synthesized 3 dataset, five Gaussian noise levels $\{0, 50\%, \ldots, 200\%\}$ are used, resulting in 5 treatments. The percentage of noise level is determined following the same technique from Section IV. Overall, through abundant replications, the randomly generated outliers or noise provides a good simulation of the real-world scenarios without tedious artificial efforts.

5) UKBench dataset: The dataset [55] contains 6376 images with 1596 groups of 4 images known to be taken of the same object but under different conditions. We select Group $1-25$, $26-50$, $51-75$, and $1-50$ forming subsets U1, U2, U3, and U4, respectively. Given the asymmetric nature of non-rigid image registration, $(4 \times 25)^2 = 10000$ and $(4 \times 50)^2 = 40000$ image pairs from the two scales are formed, respectively. All the images are of size 640×480.

6) Holiday dataset: The dataset [56] contains 1491 images, which are mainly personal holiday photos. The dataset contains 500 image groups, each of which represents a distinct scene. We first select all the groups that have four or more images, and then randomly choose images to equalize the size of each group as 4. Finally, 25 groups are randomly selected to form 10000 image pairs. All the images are down-sampled to a resolution of 640×480.

C. Results on Feature Matching

The feature matching experiments aim to test the capability of the methods on removing outliers from the given putative point correspondences. The ACF, RS, and Synthesized 1 datasets are used. RANSAC [4], PROSAC [6], LPM [14], GLPM [15], ICF [16], and GS [11] are selected for comparison. The results for ACF and RS datasets are shown in Fig. 8 and Table III, some representative examples are demonstrated in Fig 6. In order to visualize the morph of the image pairs in the examples, the motion fields are formed by first overlaying the two images, and then connecting the positions of feature points in image pairs [14], [15]. Ideally, the blue arrows (i.e., the true positive) should possess great consistency with respect to the actual image deformation. The results for Synthesized 1 dataset are shown in Fig. 7. Our SIR achieves the highest f1-score on ACF and RS datasets. For Synthesized 1 dataset, when the IR is varied, our SIR yields the highest f1-score when IR≤ 0.6, and requires around 4s to process the 6K pairs of feature points. When the IN is varied, our SIR achieves favorable performance on f1-score and time cost.

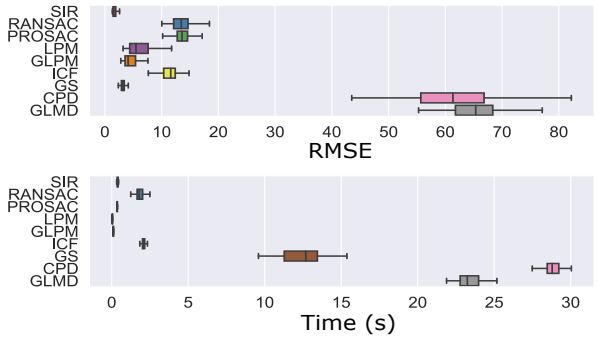

Figure 10: The boxplot of the RMSE (top) and time cost (bottom) on the Synthesized 2 dataset. 100 replications are conducted for the treatment. Best viewed in color.

D. Results on Image Registration

In the second series of experiments, we test the capability of the methods on minimizing the overlay error given an image pair. The RS, FIRE, Synthesized 2, and Synthesized 3 datasets are used. RANSAC [4], PROSAC [6], LPM [14], GLPM [15], ICF [16], GS [11], CPD [22], and GLMD [23] are selected for comparison. The results for RS and FIRE datasets are shown in Fig. 9, Fig. 14 and Table III, some representative examples are demonstrated in Fig. 13. The results for Synthesized 2 and 3 datasets are shown in Fig. 10 and Fig. 11, respectively.

For RS dataset, RANSAC and PROSAC secure a very close performance to our SIR on the satellite images, however, the inherent non-rigidity of the SUAV data prevents the parametric model based RANSAC and PROSAC from an accurate registration. The falsely matched fringe points by LPM can degrade the registration result, though they may only take

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2961480, IEEE Transactions on Image Processing

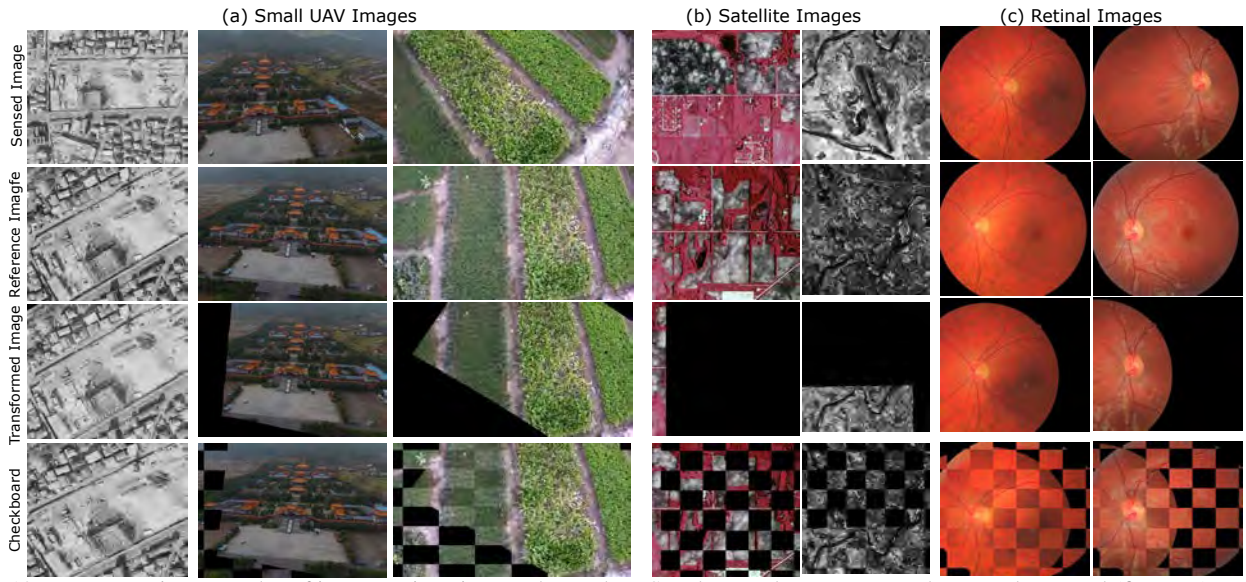

Figure 13: Representative examples of image registration results on the RS and FIRE datasets. For each group the sensed (first row), reference (second row), transformed (third row) images, and the 8×8 checkboard (last row) are displayed.

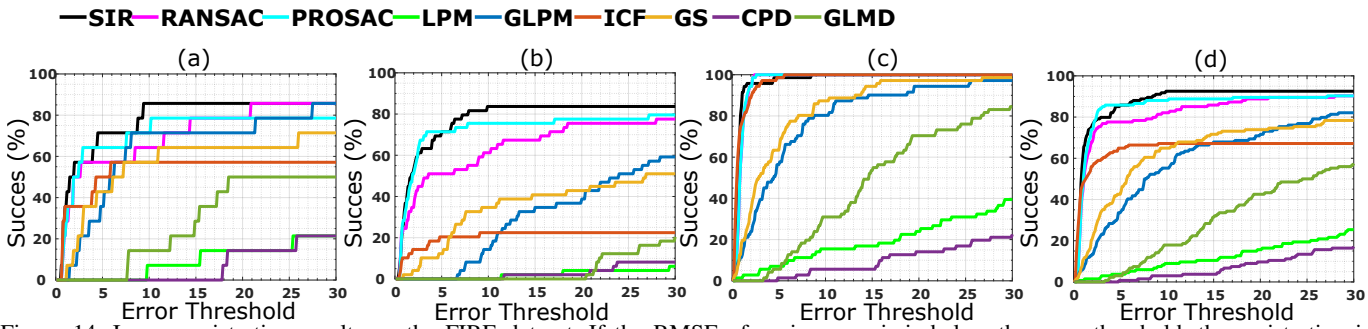

Figure 14: Image registration results on the FIRE dataset. If the RMSE of an image pair is below the error threshold, the registration is considered as successful. The success corresponds to the percentage of successfully registered image pairs for a given threshold. (a), (b) and (c): image registration results for the three categories grouped by FIRE dataset, each of which contains 14, 49, and 71 retinal image pairs. (d): the overall result for FIRE dataset. Best viewed in color. For the corresponding area under curves, please see Table III.

a very small proportion of the whole correspondence. CPD and GLMD both work unreliably as the high outlier ratio exceeds their tolerance. Our SIR achieves the lowest errors. For FIRE dataset, Our SIR acquires the largest AUC over the 134 image pairs. For Synthesized 2 dataset, under the condition of IR$=0.2$ and IN$=300$, our SIR achieves the lowest RMSE according to the boxplot, with an averagely 0.6s time cost. For Synthesized 3 dataset, Our SIR achieves the best performance for all the five noise levels.

E. Results on Image Retrieval

In the third series of experiments, we test the capability of the methods on near-duplicate image retrieval given query images. RANSAC [4], PROSAC [6], LPM [14], GLPM [15], ICF [16] and GS [11] are selected for comparison. The N-Score and total time costs are provided in Table III.

We decrease the matching strictness of our SIR by setting the CALM threshold $\lambda = 1.6$, by which our SIR gives reasonable retrieval result even when the image pairs have large depth discontinuity or motion inconsistency. Our SIR demonstrates the best N-Score. Moreover, as our SIR will break if the number of inlier pairs from the inlier pool after the pre-alignment is less than K, its average time cost has been so close to the linearithmic time featured LPM and GLPM.

VI. Discussion and Limitation

The advantages that SIR brings into the filed are as follows. Conceptually, it is a gradually generalized version of the GLPM [15] with a continuously improved dissimilarity measure. Technically, it utilizes the complement of intensity and geometric information to form a robust dissimilarity measure. It also realizes the non-parametric interpolation efficiently. Empirically, it shows favorable accuracy under non-rigid deformation, as shown in Pair 21-40 of Fig. 9, or under severe outlier degradation, i.e., inlier ratio ≤ 0.6, as shown in Fig. 7 and 10. It also shows efficiency in dealing with large feature sets, i.e., around 4s for 6000 feature pairs, as shown in Fig. 7d and 7h, or around 685s for 80K image pairs with 640×480 resolution.

The current limitation mainly lies in the dimensionality and transformation generality. First, our SIR cannot work on volumetric images (which are typically arisen in medical

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2961480, IEEE Transactions on Image Processing

Table III: The area under curve (AUC) results corresponding to Fig. 8 and 14, as well as the N-score and average time cost per image pair corresponded to image retrieval experiments. Bold fonts indicate the best results. Avg. T.: the average time cost of an image pair. "-" denotes that the method is not included in the experiment.

Figure	SIR	RANSAC	PROSAC	LPM	GLPM	ICF	GS	CPD	GLMD
Fig. 8-a	**0.8858**	0.8758	0.8514	0.8580	0.8721	0.8184	0.5460	-	-
Fig. 8-b	0.9358	0.9514	**0.9621**	0.7403	0.9276	0.7707	0.8732	-	-
Fig. 8-c	**0.9042**	0.8960	0.8917	0.7871	0.8854	0.6811	0.6541	-	-
Fig. 8-d	**0.9775**	0.8503	0.8514	0.9773	0.9619	0.8399	0.8251	-	-
Fig. 8-e	**0.9714**	0.9374	0.9621	0.7048	0.9089	0.7169	0.9156	-	-
Fig. 8-f	**0.9743**	0.8818	0.8917	0.8077	0.9343	0.6730	0.8612	-	-
Fig. 14-a	**0.8100**	0.7010	0.7126	0.0943	0.6300	0.5302	0.5590	0.0671	0.2781
Fig. 14-b	**0.7950**	0.6369	0.7235	0.0214	0.2854	0.2063	0.3473	0.0276	0.0416
Fig. 14-c	**0.9869**	0.9698	0.9700	0.1902	0.7882	0.9722	0.8349	0.0929	0.4553
Fig. 14-d	**0.8983**	0.8200	0.8530	0.1185	0.5878	0.6459	0.6278	0.0663	0.2855
U1	**2.93**	2.90	2.89	1.75	2.48	0	2.33	-	-
U2	**2.92**	2.89	2.90	1.39	2.58	0.02	1.90	-	-
U3	**2.54**	2.33	2.12	0.64	2.04	0.08	0.75	-	-
U4	**2.91**	2.86	2.84	1.64	2.38	0	2.06	-	-
H1	**2.12**	1.82	1.88	0.75	1.10	0.06	1.08	-	-
Avg. T. (s)	$1.32e-2$	$3.37e-1$	$2.40e-1$	$1.29e-2$	$\mathbf{8.57e-3}$	$1.66e+0$	$6.86e-1$	-	-

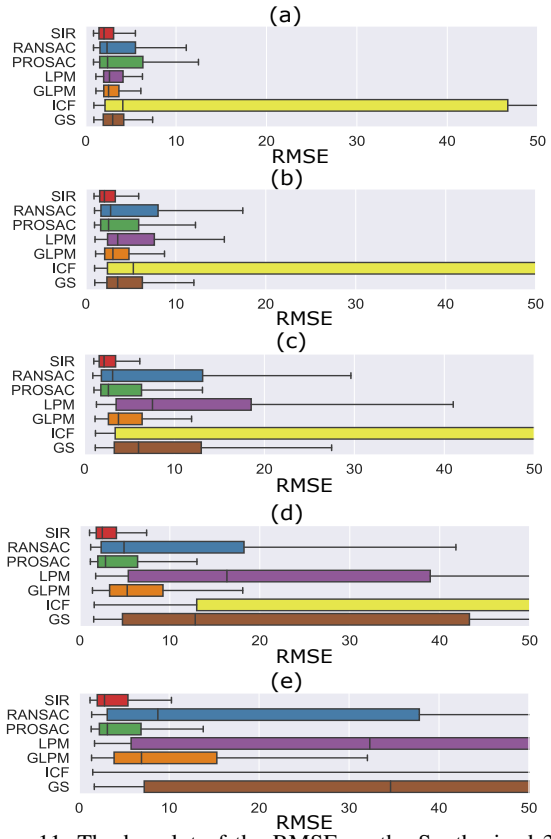

Figure 11: The boxplot of the RMSE on the Synthesized 3 dataset. From (a) to (e) are the RMSEs under $0, 50\%, \ldots, 200\%$ Gaussian noise, respectively. 100 replications are conducted for each treatment. Best viewed in color.

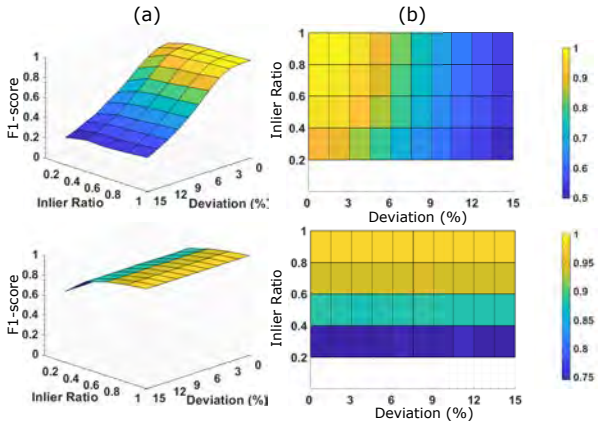

Figure 12: The f1-score on varied outlier/deviation treatments from our SIR (upper row) and GLPM (lower row). Column (a) and Column (b) are views from xyz-axis and xy-axis. 100 replications are conducted for each treatment. Best viewed in color.

image analysis), as the local histogram constituting the CALM is planar. Straightforwardly expanding the histogram to a 3-D log-polar space may produce unexpected context dissimilarity due to the high spatial sparsity. An appropriate 3-D descriptor [52] has to be designed or employed. Second, our SIR cannot work on scenarios that involve large depth discontinuity or individual motion inconsistency (which are typically arisen in stereo matching and motion tracking), due to the motion coherence [58] imposed by the ATPS. The usage of the pruning implies that the alignment of central points, or to be more general, the deformation model aligning the points, is not trivial for the CALM to work, as the error caused will be accentuated and soon makes the CALM worthless. We are currently looking at the work in [59] for ways of estimating the transformation in a patch-wise fashion, and incorporating the pruning measure ϵ as a term in the CALM.

We also attempt to explore the second limitation of our SIR quantitatively. To this end, we apply deviation to the target point set $\mathbf{Y}$ in proportion to the standard deviation produced by the normalization process. The IR and deviation strength are set to $\{0.2, 0.4, \ldots, 1.0\}$ and $\{0, 1.5\%, 3\%, \ldots, 15\%\}$, respectively. The GLPM is chosen as the baseline method. The results shown in Fig. 12 manifests that our SIR can achieve 0.8 or larger f1-score when deviation strength is $\leq 6\%$.

VII. Conclusion

We present the stepwise image registration with the closed-form solution for feature matching and image registration. Our SIR casts the image registration as a stepwise process, which gradually enriches the inlier pool and bifurcates the candidate pool. The context-aware locality measure removes mismatches with increasing accuracy thanks to the enrichment.

To avoid being trapped and time-consuming, we start from the reliable inlier pairs, and the possibility of receiving reinforcement from intermediately recovered inliers motivates us to introduce the stepwise strategy with the CALM. To reject the initially included outliers, a pre-alignment process is conducted for bifurcation before creating the pools. To maximize the finally preserved inliers, a retrieval process is included to retrieve missed outliers based on the finest CALM and alignment. The ablation study demonstrates that the stepwise manner, pruning, and retrieval process affect our SIR significantly in terms of robustness and efficiency. We have not been able to define the boundary condition of our SIR theoretically. Nevertheless, an exploratory test shows that our SIR can reach f1-score≥ 0.8 when the spatial deviation of each feature point is $\leq 6\%$ of the overall standard deviation. Experiments on feature matching, image registration, and image retrieval using real and synthesized data exhibit satisfying robustness and efficiency of our SIR against state-of-the-art methods.

VIII. Acknowledgments

The authors thank Jiayi Ma, Andriy Myronenko, and Xubo Song for providing their implementation source codes and test datasets. The authors especially thank Jiayi Ma for the proofreading and invaluable suggestions on the mathematical derivation. This greatly improved the quality of the paper. The authors also thank the volunteers from the Laboratory of Pattern Recognition and Artificial Intelligence, Yunnan Normal University for manually labeling the RS dataset. This greatly facilitated the comparison experiments. The authors also thank the associate editor and anonymous reviewers for their constructive suggestions which greatly helped improve the manuscript. This work was supported by the National Nature Science Foundation of China [41971392].

References

[1] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image Vis. Comput.*, vol. 21, no. 11, pp. 977–1000, 2003.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[3] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *MM*. ACM, 2010, pp. 1469–1472.

[4] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[5] P. H. Torr and A. Zisserman, "Mlesac: A new robust estimator with application to estimating image geometry," *Comput. Vis. Image Understand.*, vol. 78, no. 1, pp. 138–156, 2000.

[6] O. Chum and J. Matas, "Matching with prosac-progressive sample consensus," in *CVPR*, vol. 1. IEEE, 2005, pp. 220–226.

[7] J. Matas and O. Chum, "Randomized ransac with sequential probability ratio test," in *ICCV*, vol. 2. IEEE, 2005, pp. 1727–1732.

[8] O. Chum and J. Matas, "Optimal randomized ransac," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1472–1482, 2008.

[9] M. Leordeanu and M. Hebert, "A spectral technique for correspondence problems using pairwise constraints," in *ICCV*, vol. 2. IEEE, 2005, pp. 1482–1489.

[10] L. Torresani, V. Kolmogorov, and C. Rother, "Feature correspondence via graph matching: Models and global optimization," in *ECCV*. Springer, 2008, pp. 596–609.

[11] H. Liu and S. Yan, "Common visual pattern discovery via spatially coherent correspondences," in *CVPR*. IEEE, 2010, pp. 1609–1616.

[12] F. Zhou and F. De la Torre, "Deformable graph matching," in *CVPR*. IEEE, 2013, pp. 2922–2929.

[13] J. Ma, J. Zhao, H. Guo, J. Jiang, H. Zhou, and Y. Gao, "Locality preserving matching," in *IJCAI*. AAAI Press, 2017, pp. 4492–4498.

[14] J. Ma, J. Zhao, J. Jiang, H. Zhou, and X. Guo, "Locality preserving matching," *Int. J. Comput. Vis.*, pp. 1–20, 2018.

[15] J. Ma, J. Jiang, H. Zhou, J. Zhao, and X. Guo, "Guided locality preserving feature matching for remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, 2018.

[16] X. Li and Z. Hu, "Rejecting mismatches by correspondence function," *Int. J. Comput. Vis.*, vol. 89, no. 1, pp. 1–17, 2010.

[17] Y. Lipman, S. Yagev, R. Poranne, D. W. Jacobs, and R. Basri, "Feature matching with bounded distortion," *ACM Trans. Graph.*, vol. 33, no. 3, p. 26, 2014.

[18] J. Ma, J. Zhao, J. Tian, A. L. Yuille, and Z. Tu, "Robust point matching via vector field consensus." *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1706–1721, 2014.

[19] A. L. Yuille and N. M. Grzywacz, "A computational theory for the perception of coherent visual motion," *Nature*, vol. 333, no. 6168, p. 71, 1988.

[20] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, 1992.

[21] H. Chui and A. Rangarajan, "A new point matching algorithm for non-rigid registration," *Comput. Vis. Image Understand.*, vol. 89, no. 2-3, pp. 114–141, 2003.

[22] A. Myronenko and X. Song, "Point set registration: Coherent point drift," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2262–2275, 2010.

[23] Y. Yang, S. H. Ong, and K. W. C. Foong, "A robust global and local mixture distance based non-rigid point set registration," *Pattern Recognit.*, vol. 48, no. 1, pp. 156–173, 2015.

[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[25] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," in *CVPR*, 2017, pp. 6148–6157.

[26] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser, "3dmatch: Learning local geometric descriptors from rgb-d reconstructions," in *CVPR*, 2017, pp. 1802–1811.

[27] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "An unsupervised learning model for deformable medical image registration," in *CVPR*, 2018, pp. 9252–9260.

[28] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: a learning framework for deformable medical image registration," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1788–1800, 2019.

[29] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning for fast probabilistic diffeomorphic registration," in *MICCAI*. Springer, 2018, pp. 729–738.

[30] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *CVPR*, 2017, pp. 652–660.

[31] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *NIPS*, 2017, pp. 5099–5108.

[32] B. S. Hua, M. K. Tran, and S. K. Yeung, "Pointwise convolutional neural networks," in *CVPR*, 2018, pp. 984–993.

[33] S. Xie, S. Liu, Z. Chen, and Z. Tu, "Attentional shapecontextnet for point cloud recognition," in *CVPR*, 2018, pp. 4606–4615.

[34] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K. R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nat. Commun.*, vol. 10, no. 1, p. 1096, 2019.

[35] X. Pennec and J. P. Thirion, "A framework for uncertainty and validation of 3-d registration methods based on points and frames," *Int. J. Comput. Vis.*, vol. 25, no. 3, pp. 203–229, 1997.

[36] K. Teelen and P. Veelaert, "Computing the uncertainty of transformations in digital images," in *Vision Geometry XIII*, vol. 5675. International Society for Optics and Photonics, 2005, pp. 1–12.

[37] C. Stewart, "Uncertainty-driven, point-based image registration," in *Handbook of Mathematical Models in Computer Vision*. Springer, 2006, pp. 221–235.

[38] H. Mirzaalian, T. K. Lee, and G. Hamarneh, "Uncertainty-based feature learning for skin lesion matching using a high order mrf optimization framework," in *MICCAI*. Springer, 2012, pp. 98–105.

[39] T. Lotfi Mahyari, "Uncertainty in probabilistic image registration," Ph.D. dissertation, Applied Sciences: School of Computing Science, 2013.

[40] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[41] Z. Wang, B. Fan, and F. Wu, "Local intensity order pattern for feature description," in *ICCV*. IEEE, 2011, pp. 603–610.

[42] Y. T. Hu, Y. Y. Lin, H. Y. Chen, K. J. Hsu, and B. Y. Chen, "Matching images with multiple descriptors: An unsupervised approach for locally adaptive descriptor selection," *IEEE Trans. Image Process*, vol. 24, no. 12, pp. 5995–6010, 2015.

[43] Y. T. Hu and Y. Y. Lin, "Progressive feature matching with alternate descriptor selection and correspondence enrichment," in *CVPR*, 2016, pp. 346–354.

[44] J. C. Cech, J. Matas, and M. Perd'och, "Efficient sequential correspondence selection by cosegmentation," in *CVPR*. IEEE, 2008, pp. 1–8.

[45] J. Cech, J. Matas, and M. Perdoch, "Efficient sequential correspondence selection by cosegmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1568–1581, 2010.

[46] Y. Zheng and D. Doermann, "Robust point matching for nonrigid shapes by preserving local neighborhood structures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 643–649, 2006.

[47] J. Ma, J. Zhao, and A. L. Yuille, "Non-rigid point set registration by preserving global and local structures," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 53–64, 2016.

[48] S. Zhang, Y. Yang, K. Yang, Y. Luo, and S. H. Ong, "Point set registration with global-local correspondence and transformation estimation," in *ICCV*. IEEE, 2017, pp. 2669–2677.

[49] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *Int. J. Comput. Vis.*, vol. 65, no. 1-2, pp. 43–72, 2005.

[50] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, 2002.

[51] G. Donato and S. Belongie, "Approximate thin plate spline mappings," in *ECCV*. Springer, 2002, pp. 21–31.

[52] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, J. Wan, and N. M. Kwok, "A comprehensive performance evaluation of 3d local feature descriptors," *Int. J. Comput. Vis.*, vol. 116, no. 1, pp. 66–89, 2016.

[53] K. Yang, A. Pan, Y. Yang, S. Zhang, S. H. Ong, and H. Tang, "Remote sensing image registration using multiple image features," *Remote Sens.*, vol. 9, no. 6, p. 581, 2017.

[54] C. Hernandez-Matas, X. Zabulis, A. Triantafyllou, P. Anyfanti, S. Douma, and A. A. Argyros, "Fire: fundus image registration dataset," *J. Model. Ophthalmol.*, vol. 1, no. 4, pp. 16–28, 2017.

[55] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *CVPR*, vol. 2. Ieee, 2006, pp. 2161–2168.

[56] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *ECCV*. Springer, 2008, pp. 304–317.

[57] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[58] A. L. Yuille and N. M. Grzywacz, "A mathematical analysis of the motion coherence theory," *Int. J. Comput. Vis.*, vol. 3, no. 2, pp. 155–175, 1989.

[59] J. Li, Q. Hu, M. Ai, and R. Zhong, "Robust feature matching via support-line voting and affine-invariant ratios," *ISPRS J PHOTOGRAMM*, vol. 132, pp. 61–76, 2017.

Su Zhang received his bachelor degree from Xiamen University, China in 2013, and master degree from Yunnan Normal University in 2018. He is currently working towards the PhD degree in the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His current research interests include machine learning, deep learning and emotion recognition.

Wanjing Zhao received the BS degree from Yunnan Normal University, China in 2018. She is currently working towards the master degree in the School of Information Science and Technology, Yunnan Normal University. Her current research interests include image registration, point set registration and pattern recognition.

Xuying Hao is now working towards the bachelor degree in the School of Information Science and Technology, Yunnan Normal University, China. Her research interest covers point set registration and remote sensing image processing.

Yang Yang received the master degree from Waseda University, Japan, in 2007, and the PhD degree from the National University of Singapore, Singapore in 2013. He is currently a Professor with the School of Information Science and Technology, Yunnan Normal University. His research interests cover computer vision, remote sensing, and medical imaging. He was selected into the Yunnan Province Ten-Thousand Talents Program in 2018.

Cuntai Guan (S91, M92, SM03, F18) received his PhD degree from Southeast University in 1993. He is a Professor in the School of Computer Science and Engineering, Nanyang Technological University, Singapore. Prior to joining NTU, he was the founding Department Head of Neural & Biomedical Technology Department and a Principal Scientist at the Institute for Infocomm Research, Agency for Science, Technology, and Research, Singapore. He served as the founding Co-Director of the Rehabilitation Research Institute of Singapore. He serves as an Associate Editor for IEEE Transactions on Biomedical Engineering, Neurocomputing, Brain-Computer Interfaces, Frontiers in Neuroscience. His research interests are in the fields of Brain-Computer Interfaces, Neural Signal Processing, Neural Image Processing, Machine Learning, and Data Analytics. He is a recipient of the Annual BCI Research Award, the IES Prestigious Engineering Achievement Award, Achiever of the Year (Research) Award, Finalist of President Technology Award, and winner of BCI Competitions. He is a Fellow of IEEE.